AN INVESTIGATION OF THE QUALITY OF STUDENT-DEVELOPED SURVEYS AND RATING SCALES AND PSYCHOMETRIC REPORTING PRACTICES IN DOCTORAL DISSERTATIONS

By

KATHERINE HOLE
M.A., The Chicago School of Professional Psychology, 2007
B.S., Pittsburg State University, 2004

Submitted to the graduate degree program in Psychology and Research in Education and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Chairperson, Bruce Frey, Ph.D.

_____
William Skorupski, Ed.D.

_____
Vicki Peyton, Ph.D.

_____
David Hansen, Ph.D.

_____
Phil McKnight, Ph.D.

Date Defended: April 29, 2015

The Dissertation Committee for Katherine Hole
certifies that this is the approved version of the following dissertation:

AN INVESTIGATION OF THE QUALITY OF STUDENT-DEVELOPED SURVEYS AND RATING SCALES AND PSYCHOMETRIC REPORTING PRACTICES IN DOCTORAL DISSERTATIONS

_____
Chairperson, Bruce Frey, Ph.D.

Date Approved:

ABSTRACT

Doctoral dissertation research has been criticized for its quality and contribution to scholarly research. Commonalities between doctoral research and reviews of published articles indicate a lack of psychometric reporting practices of those utilizing instrumentation. Surveys, specifically those that are self-developed, have not been examined in doctoral research. Multiple sources advise researchers who create their own surveys for data collection to follow specific item writing and rating scale development guidelines.

This previous research led to the investigation of student-developed surveys and rating scales in doctoral dissertations, and the psychometric reporting practices of the students, to identify trends in graduate research. Two-hundred forty-six doctoral dissertations were examined, which included the use of 280 self-developed surveys. Specific guidelines were created for assessing the survey characteristics and students' reporting of psychometric properties.

The survey items and rating scale characteristics were considered favorable; the authors mostly adhered to survey development guidelines. The frequency of students who validated their surveys was superb. However, the lack of reliability reporting by students calls into question not only the knowledge and experience with reliability methods of the students and their committee members, but also the psychometric training students experience in graduate school.

ACKNOWLEDGEMENTS

Completing this journey would not have been possible without the support of my family, friends, and committee members. Thank you for your feedback and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Surveys are a popular method of data collection for professionals and students (Alidousti, Khosrowjerdi, Shahriari, Shirani, & Tarnoni, 2009; Fink, 2003b; Haller, 1979; Lunenburg & Irby, 2008). Differing from a test of knowledge in which there is a right or wrong answer, a survey is an instrument of data collection commonly used to gauge respondents' ratings of their first-hand perceptions or attitudes regarding some construct of interest (Spector, 1992). This mode of data collection is becoming more prominent in doctoral dissertation research. Wick and Dirkes' (1973) analysis of data collection instruments used in dissertations (n = 199) showed that 20% were using rating scales, 24% were using questionnaires, and 25% were using surveys. (The use of rating scales and questionnaires was noted to be indicative of survey research.) Survey research was being used between 13 and 48% of Adams and White's (1994) sample of 830 dissertations across six fields of study. Nelson and Coorough (1994) found in their sample of Ph.D. and Ed.D. dissertations that a descriptive research design was the most dominant, and use of frequencies or percentages was the most prevalent use of statistical analysis, evidence of a higher use of survey research. In a study of doctoral research produced in Turkey, Karadag (2011) reported scales were being used in more than half of the sample of dissertations published between 2003 and 2007.

According to Meier and Davis (1990), approximately one-third of surveys used in dissertations are developed by the doctoral investigators (i.e., created for the purpose of the study). Other studies have shown the use of any type of investigator-developed instrumentation is prevalent in dissertation research, such as in Wick and Dirkes' (1973) study in which 21% of the dissertations included tests that were created by the doctoral investigators. In a sample of

studies published across a 7-year span, Tinsley and Irelan (1989) found that of those using instruments for data collection (n = 425 instruments), up to 47% of the instruments were designed specifically for the respective studies.

With the widespread use of survey research in dissertations, and the increase of researchers developing their own instrumentation, an investigation of surveys in doctoral research is necessary as the quality of a survey can affect the overall quality of the dissertation. Research on the lack of reporting key components of survey design, implementation, and evaluation is abundant, leaving readers to question the quality of the survey, the data obtained from it, and overall interpretations and conclusions (Tinsley & Irelan, 1989). Examining this area of research in dissertations may counter or support claims such as that made by Haller (1979), who stated "The heavy reliance on these techniques, it can be argued, is another reason why doctoral dissertations are less informative than they might otherwise be" (p. 48). According to Haller (1979), the use of questionnaires in dissertations is evidence of students who hold a more "instrumental view" of (i.e., devalue) their research (p. 59). In Haller's (1979) review, 43% of the sample included questionnaire methodologies. He concluded that those students who held an instrumental view of their research did so due to choosing a questionnaire methodology for the purpose of expediency, fitting a topic to the method. Haller (1979) further concluded that students only complete the dissertations "primarily because they have to, not because they believed their research will be a contribution to knowledge in their field or because doing one is a good way to learn about the process of conducting research" (p. 60).

Doctoral Dissertations

A dissertation is completed by a doctoral student as a "rite of passage" from the role of student to professional (Haller, 1979, p. 62; Hamilton, 1993, p. 50; Lovitts, 2007, p. 29). Among

many descriptions, a dissertation is "the cumulative, tangible *best evidence* of faculty and student interest in serious and incisive scholarship" (Thompson, 1988, p. 1). The completion of a dissertation exemplifies many things (according to varying sources to account for differences within and between fields of study): mastery of a field of study; the knowledge, training, analytic and writing abilities, or technical skills and competencies, gained while in graduate school; evidence of a student's ability to conduct independent, original, or significant research; and contributing to, creating new, or verifying accepted knowledge (Alvarez, Canduela, & Raeside, 2012; Coorough & Nelson, 1997; Hamilton, 1993; Isaac, Quinlan, & Walker, 1992; Lovitts, 2007; Lunenburg & Irby, 2008; Porter, Chubin, Rossini, Boeckmann, & Connolly, 1982; Quarles & Roney, 1986; Tansey, Zanskas, & Phillips, 2012; Tewari, 2012; Thompson, 1988).

Dissertation research, however, is not increasingly contributing to, or impacting, the development of students' respective fields (Cleary, 1992, 2000; Felbinger, Holzer, & White, 1999), and is not receiving as much attention as it should (Coorough & Nelson, 1997). There is a lack of quality in doctoral dissertations, which is discouraging if the quality of dissertations can impact the reputation of doctoral programs or departments (Hamilton, Johnson, & Poudrier, 2010; Isaac et al., 1992; Thompson, 1994a). Studies and reviews have been conducted regarding the general quality of dissertations (Adams & White, 1994; Felbinger et al., 1999; Lovitts, 2007; Thompson, 1994a); to improve dissertation quality and the dissertation process (Burnett, 1999; Hamilton, 1993; Isaac et al., 1992; Lovitts, 2007; Perlmutter, 2006; Ponticell & Olivarez, 1997; Quarles & Roney, 1986; Thompson, 1987, 1994a); and to address the research design (Cleary, 1992, 2000; Coorough & Nelson, 1997; Nelson & Coorough, 1994; Tansey et al., 2012; Winter, Griffiths, & Green, 2000), measurement instruments or measurement reporting (Karadag, 2011; Thompson, 1988, 1994a), type of methodology or statistical analysis used (Adams & White,

1994; Coorough & Nelson, 1997; Isaac et al., 1992; Nelson & Coorough, 1994; Tansey et al., 2012; Thompson, 1988, 1994a; Winter et al., 2000), and results reporting or conclusions made in dissertation research (Cleary, 2000; Coorough & Nelson, 1997; Thompson, 1988, 1994a; Winter et al., 2000). The study of the surveys used in dissertation research is sparse.

Survey research in dissertations. The use of surveys as data collection tools in dissertation research has been continually increasing (Coorough & Nelson, 1997), but the question of whether the surveys used in dissertations are appropriate and quality data collection instruments has yet to be addressed. Kohr and Suydam (1970) noted various types of surveys are being used to collect data that cannot be otherwise obtained directly; "such surveys are poorly designed" (p. 78). Issues with the quality of survey research in refereed journals, which may also be evident in dissertation research, exist in the forms of defective or inappropriate research designs (Bailar & Lanphier, 1978; Fincham & Draugalis, 2013); psychometric evidence not reported (Barry, Chaney, Piazza-Gardner, & Chavarria, 2013; Bennett et al., 2011; Hogan & Agnello, 2004; Meier & Davis, 1990; Whittington, 1998) and incorrect inferences based on the results (Bailar & Lanphier, 1978). According to Draugalis, Coons, and Plaza (2008), the poor quality of survey research "can be attributed to 2 primary problems: (1) ineffective reporting of sufficiently rigorous survey research, or (2) poorly designed and/or executed survey research, regardless of the reporting quality" (p. 1).

Measurement has been defined as including a variety of elements, used as an umbrella term to refer to the levels of measurement (e.g., nominal, ordinal, interval, and ratio), test theory (i.e., classical test theory and item response theory), test construction procedures (i.e., scales, scaling), and psychometric properties (i.e., reliability and validity) of instrumentation (Aiken, West, & Millsap, 2008; Cohen & Swerdlik, 2002; Dane, 2011; Netemeyer, Bearden, & Sharma,

2003). Regarding survey research, measurement is known as the process of quantifying psychological or educational constructs that are otherwise not observable, by generating an observable response through the use of a measurement tool (Carmines & Zeller, 1979; Gliem & Gliem, 2003; Osterlind, 2010; Thorndike, 2005). Thus, a numerical representation of the degree or frequency of a particular attribute is measured, not the respondents themselves (Netemeyer et al., 2003). Reliability and validity are used as evaluations of measurement in survey research; therefore, "measurement reporting" in the current study refers to the reporting of reliability and validity evidence, used interchangeably with "psychometric reporting".

Failing to report, or inadequately reporting, measurement characteristics impacts the conclusions a reader makes about a study, such as that the study was improperly conducted (Heiman, 2001; Whittington, 1998). "The poor practice of measurement is less evident than the failure to report sufficient information for the reader to make a judgment" (Whittington, 1998, p. 33). Regardless of the type of instrumentation used for data collection, erroneous reporting of reliability and validity is evident in refereed journals and dissertations, if even addressed at all (Karadag, 2011; Qualls & Moss, 1996; Thompson, 1988, 2003). Research on the quality of measurement reporting in refereed journals is abundant; unfortunately, measurement reporting in dissertation research is deficient. Whittington (1998) stated a need for improving the quality of measurement reporting:

> Researchers build on the work of others. To the extent that the foundation of a line of studies is soft, the meaning of the body of evidence is questionable. Reviews of literature abound with discussions of inconsistent results. Is it possible that many of these inconsistencies are simply due to poor measurement? Even more serious is the impact that research has had on the practices and decisions of teachers, parents, psychologists,

> and policy makers. ...it is clear that improvement in the reporting of measurement is needed in published research. (p. 34)

Measurement reporting is a vital component of a dissertation if it is expected to contribute to the development of its particular field. Measurement is also a key piece to be reported in survey research, especially of studies that use researcher-developed surveys. "Researchers [are to] provide quantitative information about scales so that researchers, reviewers, and readers may adequately judge for themselves the strength of the study's measurement characteristics" (Meier & Davis, 1990, p. 113).

Research Purpose

Within the research on the quality of doctoral dissertations, there is yet to exist an examination of the surveys used as data collection tools. The purpose of the current study was to review student-developed surveys in doctoral dissertations, specifically to describe the characteristics reported in dissertations regarding the objectives, design, and psychometric properties specific to the use of the surveys, and to investigate the alignment of these to best practice. The type of surveys to be addressed included those that were self-administered and used rating scales (e.g., Likert-type) (see Likert, 1932); these are also known as "social surveys" and are commonly used in dissertation research (Alidousti et al., 2009; Benson & Clark, 1982; Fink, 2003b; Haller, 1979; Lunenburg & Irby, 2008). Three research questions were addressed:

1. What is the quality of student-developed survey items in doctoral dissertations?
2. What is the quality of the rating scales associated with student-developed surveys in doctoral dissertations?
3. What are the reporting practices of psychometric properties for student-developed surveys in doctoral dissertations?

Because “the quality of measurement in survey research varies” (Fowler, 1995, p. 4), a study of the quality, psychometric reporting, and other characteristics of student-developed surveys can bring forth the issues surrounding this popular type of data collection used in doctoral research. Detailing trends now could remedy issues in future doctoral surveys, which can impact the overall quality of a dissertation and future publications.

Summary

Surveys and associated measurement reporting are areas that are scarcely investigated in doctoral dissertation research. The research purpose of the current study was to assess the quality of student-developed surveys in doctoral dissertations, as well as to investigate psychometric reporting practices. The following literature review pertains to the quality of dissertations. The elements of constructing a survey, including its design, implementation, and evaluation, are further addressed to discuss what is a quality survey and what steps a doctoral student should take to create one.

CHAPTER TWO

REVIEW OF THE LITERATURE

The purpose of graduating students from doctoral programs is so those students will continue their research and teaching as professionals. Winter et al. (2000) discussed the root of the term "doctorate" (*docere*), which means "to teach" (p. 36), connecting doctoral work (specifically the dissertation) to its root meaning, acknowledging that "teaching" others could be in the form of publishing in refereed journals. According to Felbinger et al. (1999), "the traditional view of doctoral education…is the reproduction of the professoriate to ensure continued knowledge development through research and the dissemination of knowledge through teaching" (p. 459).

Spriestersbach and Henry (1978) interrogated the role of the dissertation in doctoral programs, stating burdens such as students being bound by their topics, or that students who move into faculty roles were not likely to continue studies related to their dissertation research. The authors questioned whether the current use of the dissertation was necessarily the best, and recommended that the notion of the dissertation as a "significant contribution to knowledge" be buried. Others have noted that when a student contributes to a field of study, the student will feel a sense of belonging as a scholar within that field (Finney & Pastor, 2012; Kamler & Thomson, 2008). The dissertation has been credited by "All But Dissertation" (ABD) students and Doctor of Philosophy (Ph.D.) graduates in research and non-research occupations as favorable and valuable, emphasizing the degree obtained is "a valued credential" (Porter et al., 1982, p. 481).

Spriestersbach and Henry (1978) recommended programs continue to evaluate their criterion measures of a doctoral education, perhaps a student should be offered an individualized program based on the student's history and experience, and changes be made to the dissertation

requirements. The dissertation should be part of the research experience while in graduate school, not the ultimate obstacle over which a student has to jump (Porter et al., 1982; Spriestersbach & Henry, 1978). Similarly, the dissertation has been described as being "an integral part of doctoral education rather than an exit outcome...Graduate programs in education should develop specific, commonly understood, explicit indicators of quality in education dissertations based on a level of mastery expected of novice researchers" (Ponticell & Olivarez, 1997, p. 121), similar to how these institutions discuss "performance indicators for measuring quality" (Harvey & Green, 1993, p. 25).

Lovitts (2007) also addressed the need for explicit indicators of dissertation quality. Lovitts' (2007) study of focus group data from multiple faculty members across a variety of universities unveiled that some faculty members viewed the dissertation as an outcome-based assessment, similar to Cleary (1992) and Porter et al. (1982) who characterized the dissertation as a capstone experience in graduate education. "This attention to outcomes is part of a larger discussion in doctoral education about how best to prepare doctoral students for the professional (academic and nonacademic) destinations" (Lovitts, 2007, p. 22). Stressing the need for performance expectations of a dissertation within graduate programs so the faculty, department, and discipline's expectations are explicit, Lovitts (2007) begged the question: "Should the evaluation focus exclusively on the product or should knowledge about the person and the process be factored into the equation?" (p. 24). Hamilton et al. (2010) argued the evaluation of a dissertation should be context-specific, noting students have varying backgrounds, experiences, critical thinking skills, writing skills, verbal abilities, among other differences, which may affect the quality of the dissertation. Regardless, the concrete evidence of obtaining a terminal degree should be flawless. According to Adams and White (1994), "dissertation research should not

have obvious flaws. It should be directed by an advisor and a committee that are concerned enough with the quality of the work to correct the most obvious flaws" (p. 567). The roles of the advisor and committee are essential to dissertation research (Burnett, 1999): according to Kamler and Thomson (2008), advisors should pay more attention to doctoral writing so that students do not fall back on inadequate resources; and according to Lovitts (2007), "the only quality rating of dissertations currently available at most American universities is the recommendation of the dissertation committee" (p. 3). Porter et al. (1982) reported that doctoral students perceived supervision and evaluations of dissertation research as important by their supervisors and committees. A positive dissertation experience was related to a close relationship between the student and supervisor, acknowledging "the dissertation teaches respect for the scientific method in a way that nothing else could" (Porter et al., 1982, p. 481). Felbinger et al. (1999) associated dissertation-to-advisor ratios with the "poor quality of…dissertation research" (p. 461).

In his informal review of social science dissertations, Perlmutter (2006) identified characteristics that were becoming too common, stating the studies were too short; one-note analyses (e.g., the use of one population or one set of research questions); unoriginal; and poorly written. He claimed these characteristics stemmed from the student's rush to finish, inquiring "Do we surrender to expediency or hold fast to standards of quality?" (Perlmutter, 2006, para. 5). But what constitutes a 'quality' dissertation? In higher education, like other businesses or institutions, "quality matters" (Harvey & Green, 1993, p. 9). The concept of quality, however, is relative and depends on who is using the term and "the circumstances in which it is invoked" (Harvey & Green, 1993, p. 10). Regarding the dissertation, quality most likely depends on the perspectives of committee members. "Quality can be viewed as exceptional [i.e., exceeding high standards or passing minimum standards], as perfection (or consistency) [set specifications], as

fitness for purpose, as value for money and as transformative" (Harvey & Green, 1993, p. 11). Although Tewari (2012) concluded that the American dissertation process was more rigorous and had more quality control mechanisms than other models, dissertations in the United States and abroad have been scrutinized for their quality (Cleary, 1992, 2000; Felbinger et al., 1999; Lovitts, 2007; Perlmutter, 2006).

Perlmutter (2006) addressed doctoral students' lack of individual scholarly work, stating a dissertation "needs to be a rich, multifaceted document that can produce a considerable body of published scholarship" (para. 8). The quality of 'published scholarship,' however, has been called into question for over 50 years. If research is a "delivery system" by which professionals depend on and acquire much of their understanding of educational problems (Hall, Ward, & Comer, 1988; Ward, Hall, & Schramm, 1975) and "requires professional objectivity and integrity" (Ponticell & Olivarez, 1997, p. 114), the number of studies conducted on the quality of published research should not be as abundant as it is (e.g., Goodwin & Goodwin, 1985a, 1985b; Hall et al., 1988; Hastings & Stewart, 1983; Kieffer, Reese, & Thompson, 2001; Kupfersmid, 1988; Meier & Davis, 1990; Ruja, 1955; Thompson, 1994c; Tuckman, 1990; Vockell & Asher, 1974; Ward et al., 1975; West, Carmody, & Stallings, 1983; Willson, 1980). Within published literature, measurement reporting has been scrutinized for its quality (e.g., Bailar & Lanphier, 1978; Barry et al., 2013; Hogan & Agnello, 2004; Kohr & Suydam, 1970; Qualls & Moss, 1996; Schaeffer & Dykema, 2011; Tinsley & Irelan, 1989; Whittington, 1998). Evaluating the components of what constitutes a quality survey, specifically reviewing resources designed to aid students and professionals alike with the design, implementation, and evaluation of a survey, as well as reviewing studies that used surveys as data collection tools, or reviewing publications

conducted to examine survey research within other publications, may reveal the type of problems that could be associated with the use of survey research in doctoral dissertations.

Survey Research

A survey is an instrument of data collection used in many fields of study, described as "a research method by which information is gathered by asking people questions on a specific topic and the data collection procedure is standardized and well defined" (Bennett et al., 2011, p. 3). Surveys are commonly used to quantify human constructs that are otherwise not observable (i.e., subjective); such surveys are used to collect respondents' information to compare, describe, explain, or count their attitudes, behaviors, expectations, feelings, judgments, knowledge, opinions, perceptions, personalities, traits, or values (Alvarez et al., 2012; Barry et al., 2013; Creswell, 2014; Dawis, 1987; Draugalis et al., 2008; Fink, 2003a, 2003d; Fowler, 1995; Groves, 2011; Haller, 1979; Heiman, 2001; Likert, 1932; Netemeyer et al., 2003; Spector, 1992; Thorndike, 2005). Hastings and Stewart (1983) stressed that "in order for the reader to have assurance that the research was adequately executed and that the test results were significant, there must be a clear explanation of test construction procedures" (p. 702). A test is a collection of items that measure a respondent's knowledge of content of interest. Test construction, as described, can be translated to survey development, or the process by which the researcher designs, implements, and evaluates a survey for data collection.

When an appropriate survey is not available to collect data for the purposes of a study, a new instrument can be developed by the researcher. Regardless if the survey is new, modified, or borrowed, "Researchers should follow standard and systematic methods that aim to improve the quality of measuring instruments" (Coluci, 2012, p. 121). A researcher who develops a survey for data collection should fully provide how it was developed (Draugalis et al., 2008).

Fincham and Draugalis (2013) stressed the need for survey research guidelines, as "poor reporting guidelines lead to subsequent deficient outcome segments in written summaries of research" (p. 2). According to Fowler (1995),

> ... cognitive testing, good field pretests, and appropriate validating analyses provide scientific, replicable, and quantified standards by which the success of question design efforts can be measured...there is no excuse for question design to be treated as an artistic endeavor. Rather, it should be treated as a science. Unfortunately, there is a long history of researchers designing questions, in a haphazard way, that do not meet adequate standards. (p. 154)

Of interest in the current study is the method of self-administration, in which the respondent self-reports (Netemeyer et al., 2003), as this is commonly used in dissertations due to ease of administration. According to Coluci (2012) a disadvantage to self-administered surveys "is that the researcher can not [sic] clarify any doubt of the subjects, even if the researcher is present. The advantage is that there is less bias to answer, i.e., less interference from the researcher in the subject's response" (p. 127). In regards to subject matter that is considered socially undesirable, self-administration has been found to produce higher reports, as this mode of data collection is private (Schaeffer & Presser, 2003).

Noted as a popular method of data collection in doctoral dissertation research, rating scales (i.e., Likert-type) are commonly used in self-report surveys, take less time to construct and complete, and are easier to administer (Benson & Clark, 1982; Converse & Presser, 1986; Heiman, 2001; Likert, 1932; Krosnick, 1999; Litwin, 2003; Saris, Revilla, Krosnick, & Shaeffer, 2010; Spector, 1992). Across survey research literature there is consensus regarding the steps to be taken to design, implement, and evaluate a quality survey and the data obtained from it. The

following is a description of this consensus, which focuses on the use of rating scales to collect data.

Developing a Quality Survey

Checklists for survey research have been created, aiding researchers in developing their instruments by following specific guidelines. Such components of a survey include addressing the content of the survey and the appropriateness of using a survey design in the study; identifying the process of survey development; identifying the procedure for pilot-testing the survey; identifying the variables of the study, and aligning those variables to the research questions and survey items; establishing evidence of reliability and validity; and reporting statistical analyses to be used to address the research questions (Creswell, 2014; Fink, 2003a, 2003d; Lunenburg & Irby, 2008; Netemeyer et al., 2003). Some of the characteristics of the best surveys include specific, measurable objectives; straightforward items; and evidence of reliability and validity (Fink, 2003a, 2003d; Fowler, 1995; Spector, 1992). The following review of these topics includes an integration of research on best practices, including an assessment of the lack of reporting of psychometric components in doctoral dissertations.

Survey blueprint. The first step prior to designing a survey is to create a blueprint by specifying the survey's purpose, objectives, and construct(s) of interest (Benson & Clark, 1982; Fowler, 1995, 2009; Spector, 1992). The survey blueprint is a means of linking the purpose of the survey to the study. The purpose of the survey, the objectives of it, as well as the constructs to be measured by it, are all included in a blueprint. "A good survey instrument must be custom made to address a specific set of research goals" (Fowler, 1995, p. 78). The need for or significance of the survey, the research design of the study, and the overall plan for conducting the study should be included (Fink, 2003d).

An objective of a survey is "a statement of the survey's hoped-for outcomes" (Fink, 2003d, p. 6), and can include the overall purpose of the survey or the definitions of the constructs to be measured by the survey (Benson & Clark, 1982; Coluci, 2012). Objectives of the survey can be identified through the purpose of the study, because "surveys must be custom-built to the specifications of given research purposes" (Converse & Presser, 1986, p. 7).

Typically considered the most difficult, but the most essential, step in the survey design process is "Clearly defining the construct—its facets and domains" (Netemeyer et al., 2003, p. 89). The construct's definition is to be un-confounded and appropriately represented (not too broadly or narrowly defined) (Netemeyer et al., 2003). Constructs used in survey research are typically latent, or "not directly observable or quantifiable" (Netemeyer et al., 2003, p. 7); therefore, a literature review is the first step in defining the construct(s) (Netemeyer et al., 2003; Spector, 1992). A literature review can reveal if other researchers have attempted to define the construct or create an instrument to measure that construct (Netemeyer et al., 2003). Experts of the content of interest, focus groups, and consensus panels can be employed to operationalize the construct(s) (Fink, 2003d; Fowler, 1995, 2009; Netemeyer et al., 2003; Spector, 1992).

A recommended approach to developing a survey is an inductive or confirmatory one, in which the survey is developed after defining the construct(s) (Spector, 1992). According to Spector (1992),

> It almost goes without saying that a scale cannot be developed to measure a construct unless the nature of that construct is clearly delineated… Without a well-defined construct, it is difficult to write good items and to derive hypotheses for validation purposes. (p. 12)

A deductive or exploratory approach, such as using factor analysis, occurs when an investigator creates and administers survey items, thereafter discerning the construct(s) measured (Spector, 1992).

Designing a quality survey blueprint aids the investigator in the development of the survey items and measurement scale, as well as the study's research design, research questions, and hypotheses, all of which should be linked (Coluci, 2012; Creswell, 2014; Dawis, 1987). "The specific research questions established are the guide to the specific questions or items that must be included in the survey" (Fink, 2003c, p. 6). Identifying survey objectives will ensure the data obtained from the survey are useful (Ho, 2005). Clearly defining the construct of interest aids in developing the survey items and measurement scale (Spector, 1992), and enhances the psychometric properties of survey data: "When a construct is not defined carefully in advance, there is considerable risk that the scale will have poor reliability and doubtful validity" (Spector, 1992, p. 13).

Item writing. Survey items are written to "define or operationalize the construct" (Fink, 2003b, p. 42) identified in the survey blueprint. Dawis (1987) and Netemeyer et al. (2003) recommended conducting interviews with a representative group from the target population to generate items by "elicit[ing] a wide range of statements about the variable in question" (Dawis, 1987, pp. 481-482). Conducting a focus group discussion with five to eight individuals with experiences relative to the topic can provide the researcher with valuable information regarding how the population interprets the content and to eliminate ambiguities in how items are written (Fowler, 1995). The investigator can create item stems that are easier to understand by respondents based on interviews or focus group discussions, as they would be written in a language that is common amongst the target population.

The researcher is to ensure the items written "are derived from and appropriate to the objectives of the instrument" (Benson & Clark, 1982, p. 792). A sufficient amount of items should be written to adequately cover the construct(s) to be measured (Dawis, 1987; Kitchenham & Pfleeger, 2002; Likert, 1932), which can increase the reliability of scores (Netemeyer et al., 2003; Spector, 1992). It is recommended that two times as many items initially be written than what will be in the final survey (Benson & Clark, 1982; Netemeyer et al., 2003); however, there is no set criterion of the number of items that should be included in a final survey (Netemeyer et al., 2003).

New items are written for the survey, or existing items can be modified; however, it cannot be guaranteed that reused items were good in the first place (Converse & Presser, 1986). Existing survey items should be evaluated regardless of how often they have been used (Fowler, 1995). When using or modifying existing items, it is important to confirm the purpose of the original survey is comparable to that of the investigator's research purpose. The items that are written or obtained from another source should be directly related to the objectives of the survey (Fink, 2003a). The goal of survey items is to produce quality data: "a good question is one that produces answers that are reliable and valid measures of something we want to describe" (Fowler, 1995, p. 2).

Although surveys are typically comprised of closed items (Creswell, 2014), item responses also can be open. Open-response options allow the respondent to use their own words, whereas closed-response options use preselected responses determined by the researcher (Fink, 2003d; Fowler, 1995, 2009; Krosnick, 1999). According to Fink (2003d), the results obtained from closed-response items are more likely to be reliable over time, and "lend themselves more readily to statistical analysis and interpretation" (Fink, 2003d, p. 18). Alvarez et al. (2012)

recommended that open items be minimized, making use of pre-coded response options. Reponses to open items have to be cataloged and interpreted, and can be difficult to compare or analyze (Schaeffer & Presser, 2003). However, open items are appropriate if "the range of possible answers greatly exceeds what reasonably could be provided" by the researcher or if "the answers are virtually impossible to reduce to a few words" (Fowler, 1995, p. 177). Closed items are more specific and offer some frame of reference for respondents, and if appropriately designed, the response categories help build distinctions by forcing a choice by respondents (Converse & Presser, 1986; Fowler, 1995). Whether closed items are more valid than open items is controversial (Converse & Presser, 1986). However, when using a rating scale, closed items are more reliable (Fowler, 2009).

The type of measurement scale used in a survey depends on the type of items created (Coluci, 2012): "the goal of standardized measurement is central to survey research" (Converse & Presser, 1986, p. 31). Surveys in doctoral research typically consist of closed-ended items in which a respondent rates their attitude towards the content on some type of scale (Likert, 1932). These types of survey items have two parts: the stem and the rating scale.

Item stem. The type of item stem for a rating scale is typically indirect and is a request of the respondent's opinion (Saris & Gallhofer, 2007). "It is essential that all statements be expressions of *desired behavior* and not statements of *fact*. Two persons with decidedly different attitudes may, nevertheless, agree on questions of fact" (Likert, 1932, p. 44).

Survey items are to be neutral, purposeful, straightforward, specific and concrete, mutually exclusive (identify only one idea at a time), and seem fair to the respondent (be unbiased); as well as be written in complete sentences with correct grammar and syntax (Benson & Clark, 1982; Converse & Presser, 1986; Fink, 2003d; Fowler, 1992, 1995; Kitchenham &

Pfleeger, 2002; Likert, 1932; Netemeyer et al., 2003; Spector, 1992). Requesting input from subject matter experts on both sides of a topic will aid in the avoidance of biased items (Schuman, 1986).

Item "wording should be appropriate, specific and precise, avoiding leading and loaded terms or descriptions" (Ho, 2005, p. 246). Items should have shared meaning, or be written in a way that they have the same meaning to the researcher(s) as they do the respondents, and that the understanding of the items is constant across the respondents (Converse & Presser, 1986; Fowler, 1992, 1995, 2009; Heiman, 2001; Porter, 2011). Consistency in the understanding of a survey item can improve the validity of responses (Fowler, 1995). When necessary, common concepts in Standard English should be used; definitions of uncommon concepts should be provided (Converse & Presser, 1986; Fowler, 1995; Spector, 1992). If definitions are included in the item to clarify the question, it should be provided first to ensure respondents understand the content and to evoke more valid responses (Fowler, 1995). "Definitions of key survey concepts should be based on the best available theory and practice" (Fink, 2003c, p. 26); however, "respondents cannot be expected to learn complex new material in the course of a survey question" (Converse & Presser, 1986, p. 25). The length and language of a survey item should be considered (Benson & Clark, 1982; Converse & Presser, 1986).

Items that are specific are considered better than those that are vague, as the more general an item is, the more susceptible it is to having a wider range of respondent interpretations (Converse & Presser, 1986; Heiman, 2001). In the case of rating scale items, specifically-worded items are better predictors of behavior (Converse & Presser, 1986). A survey item should only include one concept; those that address multiple concepts are considered double-barreled and should be avoided (Converse & Presser, 1986; Fowler, 1995, 2009; Heiman, 2001;

Likert, 1932; Netemeyer et al., 2003; Spector, 1992). It is difficult to gauge a respondent's true response when an item includes more than one idea (Spector, 1992).

Words in the item stem that can be considered restrictive should be avoided; positive statements are typically used for those items with a rating scale (Converse & Presser, 1986). Spector (1992) recommended the use of both positively- and negatively-worded items to reduce bias and acquiescence of responses; however, this excludes the use of negative words or terms (e.g., not) in the item stems. Respondents may misread the item by missing the negative word, thereby responding to the item on the wrong end of the scale (Spector, 1992). This can lead to reduced reliability (Netemeyer et al., 2003; Spector, 1992).

Items that are ambiguous or unclear can affect the quality of the survey data, lead to more error or inconsistency in measurement, produce biased estimates, and lead to conclusions about the results that are not valid (Fowler, 1992, 1995; Spector, 1992). Fowler's (1992) study of the effect of unclear terms on survey data revealed that revising items for clarity led to better items overall. Fowler (1992) found "revised questions that clarify and define key concepts produced significantly different distributions of answers is compelling evidence that there was considerable error in the responses to the original questions, and error was almost certainly reduced by the revised wording" (p. 228).

Small wording changes can affect or shift responses (Fowler, 1995); such effects are difficult to predict in advance and "also indicate the importance of not basing conclusions on results from a single question" (Converse & Presser, 1986, p. 42). However, small wording changes do not affect responses unless the meaning of the question itself is changed (Schuman, 1986). Converse and Presser (1986) recommended certain measures be built in to surveys: use of split-sample comparisons, such as creating item skip patterns; follow closed questions with

open questions to "provide valuable guidance in the analysis of the closed questions" (p. 43); and include multiple questions on the same topic, as it is "difficult to uncover complexity" with a single item (p. 45). Likert (1932) recommended:

> To avoid any space error or any tendency to a stereo-typed response it seems desirable to have the different statements so worded that about one-half of them have one end of the attitude continuum corresponding to the *left* or *upper* part of the reaction alternatives and the other half have the same end of the attitude continuum corresponding to the *right* or *lower* part of the reaction alternatives… These two kinds of statements ought to be distributed throughout the attitude test in a chance or haphazard manner. (p. 46)

The composition of survey items can affect measurement, validity, and reliability (Schaeffer & Dykema, 2011). Item wording, the context of the item, the rating scale, and the technique of data collection can individually and together lead to different errors and influence the type of data collected (Coluci, 2012; Fowler, 1995; Saris, Van Wijk, & Scherpenzeel, 1998). When developing survey items, researchers should take into consideration not only the wording of the item, but the order of the items in the survey, and the type of response options in the rating scale.

Rating scale. The scale of interest in the current study is the rating scale. A rating scale comprises some level of agreement (i.e., *Strongly Agree* to *Strongly Disagree*), frequency (i.e., *Never* to *Always*), importance (i.e., *Not at all Important* to *Extremely Important*), or other desired anchor (Dawis, 1987; Heiman, 2001; Litwin, 2003; Ostini & Nering, 2006; Spector, 1992). (For additional anchor options, see Vagias, 2006). The response options of a rating scale should be "exhaustive, unbiased and mutually exclusive" (Kitchenham & Pfleeger, 2002, p. 21). To reduce error in measurement that can occur when respondents interpret the scales differently, clearly

ordering the response options along a continuum is recommended (Fowler, 1995, 2009; Spector, 1992).

A rating scale can have three to 11 predefined response options (Fowler, 1995; Netemeyer et al., 2003; Saris et al., 2010; Spector, 1992). Two response options is considered a dichotomous scale; a scale with three or more response options is considered multichotomous (Netemeyer et al., 2003), which is of interest in the current study. An optimal number of categories is five to seven (Fowler, 1995; Netemeyer et al., 2003), as respondents do not provide any more meaningful information if more categories are available (Fowler, 1995). Including more response options increases the reliability of the data (Dawis, 1987) and distributes respondents across the continuum to obtain more valid information (Fowler, 1995); however, too many response options can impair reliability and validity (Netemeyer et al., 2003; Saris & Gallhofer, 2007) or the overall quality of the survey (Saris et al., 2010).

Rating scales can have a unipolar or bipolar scale (Spector, 1992). A unipolar scale is one-sided with positive (zero to positive values) or negative (zero to negative values) dimensions (Dawis, 1987; Fowler, 1995; Netemeyer et al., 2003; Spector, 1992). A bipolar scale is two-sided with negative to positive values (Dawis, 1987; Fowler, 1995; Netemeyer et al., 2003; Spector, 1992). Starting the scale with 0 versus 1 can influence the type of responses, as respondents' interpretations of the scale will differ (Schaeffer & Presser, 2003). Response options should be ordered from low to high (Spector, 1992); it has been shown that presenting the scale with the negative categories first produces more responses on that end of the scale (Fowler, 1995).

Rating scales include only values, only anchors (i.e., adjectives), or a combination of the two in which labels are assigned to each value in the scale, or at each end of the scale

(Netemeyer et al., 2003). Words, phrases, or extended prose are recommended to use as anchors (Dawis, 1987). The use of labels (not complete sentences) should add value to the numbers (Saris & Gallhofer, 2007) and not contradict, as including the labels can increase the validity of responses by dividing the scale into equal parts (Krosnick, 1999). Labeling each scale point will increase reliability (Saris & Gallhofer, 2007) and validity, as the meaning of each is therefore clarified for the respondent (Krosnick, 1999; Schaeffer & Presser, 2003). The anchors of a two-sided scale should be symmetrical on both sides of the middle point for added consistency of responses (Fink, 2003d; Fowler, 1995; Spector, 1992).

Converse and Presser (1986) recommended omitting the middle alternative (e.g., *Don't Know*, *Neither Agree nor Disagree*; *Neutral*, *No Opinion*) to measure intensity; however, there is disagreement amongst researchers as to whether to include a middle option on a rating scale. The middle option is one that is appropriate for those respondents who do not have the knowledge or opinion of the topic matter (Fowler, 1995). Dawis (1987) suggested that the middle scale point could be eliminated to make the underlying construct linear. If only the scale's extremes are anchored, the middle option should also be labeled to provide more information to the respondents and therefore increase reliability (Fowler, 1995; Netemeyer et al., 2003).

Likert (1932) initially suggested the inclusion of a middle option would not affect a respondent's reaction towards the item content. The difference between those respondents who select an intensity response versus a middle option, if given, is not affected if the middle option is not available (Converse & Presser, 1986). Respondents who feel strongly about an issue are typically not affected by a middle option, and are more likely to respond to the item (Converse & Presser, 1986; Heiman, 2001; Krosnick, 1999). Respondents who "satisfice" or expand less

effort responding to an item may be attracted to a middle option such as *Neutral* or *Don't Know* (Krosnick, 1999). The same can be said about including a *No Opinion* option: those respondents who feel strongly about a topic will not be affected by its presence (Krosnick, 1999).

> The more often a person performs behaviors that can be informed or shaped by an attitude, the more motivated that person is to form such an attitude, and the less likely that person is to say he or she has no opinion on an issue. (Krosnick, 1999, p. 557)

The use of a *No Opinion* option can increase the validity of data, as including it may deter respondents "from offering meaningless opinions" (Krosnick, 1999, p. 558). Those respondents who do not understand the question or are uncertain about the topic will more likely select this option (Krosnick, 1999). Schaeffer and Presser (2003) noted the inclusion of this option does not affect validity; however, Saris and Gallhofer's (2007) review of survey items showed that including a middle option improved the reliability and validity of the data.

Converse and Presser (1986) recommended that if including a middle option, the investigator should follow-up with a higher-intensity item to separate whether the respondents feel strongly towards an issue (i.e., a forced-choice item). A forced-choice response item is more likely to elicit a well-thought-out response than is a rating scale item, as there is a "tendency of respondents to agree irrespective of item content" (Converse & Presser, 1986, p. 38). There is greater incidence of this occurring, known as acquiescence, by those respondents who are less educated (Converse & Presser, 1986; Krosnick, 1999; Schuman & Presser, 1980). Respondents who are less educated typically will produce data that is of lower quality (Saris & Gallhofer, 2007). However, Schuman and Presser (1980) found differing results: in some of their studies, the frequency of choosing the middle option *increased* with education.

Concerns with how items are written, the appropriateness of the scale, and whether both are appropriate measures of the construct(s) of interest can be thwarted through a pilot test (Draugalis et al., 2008; Netemeyer et al., 2003). Conducting a pilot test, also known as a field test or pretesting, is a main phase of survey development, taking place after survey items are written (Fowler, 1992; Litwin, 2003).

Pilot test. Conducting a pilot test of survey items is a method of identifying potential problems with the design of the survey, or the research itself (Litwin, 2003; Sanchez, 1992). "Careful examination and pretesting of questions...can greatly improve the quality and efficiency of survey measurement" (Fowler, 1995, pp. 151-152). The survey and its items are pretested for the purposes of evaluating the "variation, meaning, task difficulty [of, and] respondent interest and attention to" an item (Converse & Presser, 1986, p. 54). Items that are borrowed from other sources must also be pretested, as the meaning of the items can be affected by their order within a survey, or by the context in which they are used (Converse & Presser, 1986). A pilot test can also be used as an initial assessment of validity, as well as to decrease the initial pool of survey items (Netemeyer et al., 2003). The results of a pilot test should be included in a study (Fowler, 2009). "Good question and instrument evaluation prior to actually doing a survey is a critical part of good survey practice. It is one of the least expensive ways to reduce error in survey estimates" (Fowler, 2009, p. 126).

Pretesting is a process that can take place multiple times, as it should be based on the query "pretested for what?" (Converse & Presser, 1986, p. 54). Two pretest phases are recommended, the first for further development of the items, followed by evaluation and a second "polishing" pretest (Converse & Presser, 1986, p. 65; Fowler, 1995). "Although rigorous, routine testing is necessary to advance survey science, better questions and better

measurements result whenever researchers take steps to critically evaluate how consistently people can understand and answer their questions" (Fowler, 1995, p. 153). The pretest phase also allows respondents to assess the overall format of the survey, including its language, type size, appropriateness, and estimated time for completion (Dawis, 1987; Litwin, 2003), as well as to check the response options and allow the researcher to evaluate reliability and validity (Dawis, 1987; Draugalis et al., 2008; Kitchenham & Pfleeger, 2002). Overall, the survey should be easy to go through; its instructions, items, rating scale, and response options should be easy to read and understand for all respondents (Fowler, 1995; Spector, 1992).

Typically a sample of respondents is drawn from the target population of interest (those who would be included in the intended sample group), although experts can be solicited to review the items for possible bias (Benson & Clark, 1982; Coluci, 2012; Converse & Presser, 1986; Fowler, 1995, 2009; Spector, 1992). "The questionnaires should be tested on prospective respondents before they are finalized for use in surveys" (Ho, 2005, p. 246). The pilot test group, recommended between 25 to 100 respondents, is utilized to report on problems with the survey, such as the clarity of item wording and ease of use, or problems with the survey format or measurement scales, through the use of focus or discussion groups (Benson & Clark, 1982; Coluci, 2012; Converse & Presser, 1986; Creswell, 2014; Dawis, 1987; Draugalis et al., 2008; Fowler, 1995, 2009; Krosnick, 1999; Litwin, 2003). Respondents from the general public are not recommended to be included in a pilot test group as they would not be as critical of, or sophisticated in, their review of the survey items (Converse & Presser, 1986). Modifications are made to the survey items after receiving feedback from the group (Coluci, 2012; Fowler, 1995), as "it is often easier to improve a bad item than to develop a new one" (Benson & Clark, 1982, p. 796).

Assessing the variation of responses within a target population is a common goal of a pilot test, as the distribution of responses is meaningful "in detecting subgroups of people or clusters of attitudes of analytical interest" (Converse & Presser, 1986, p. 55). Evaluating respondents' interpretations of an item aids in identifying potential underlying wording effects, as respondents may transform an item into something they can more easily understand and answer (Converse & Presser, 1986). The goal of this purpose of the pilot test is to ensure the respondent's meaning of the item is the same as that of the investigator (Converse & Presser, 1986). The purpose of evaluating the task difficulty of an item is similar to that of the purpose of evaluating the meaning of an item: wording effects. Respondents may not fully comprehend the item being posed, and essentially make the item more difficult to answer than originally intended (Converse & Presser, 1986).

Modifications from individual item pretesting must be addressed before the survey as a whole can be pilot tested, as the individual item changes will affect pretest results (Converse & Presser, 1986). The total survey is pretested for the purposes of assessing the "flow" and naturalness of the survey; the order of survey items; skip patterns (such as, "If you answered yes to the previous item, please skip to item 10"), although these should be minimized in self-administered surveys (Fowler, 1995); time to complete the survey; respondent interest and attention during the entire administration of the survey; and respondent well-being if the survey contains items that include sensitive subject matter (Converse & Presser, 1986; Kitchenham & Pfleeger, 2002; Litwin, 2003).

Developing a survey for data collection without first pilot-testing it can lead to problems with the survey and research design. "Researchers who do not adequately test respondent understanding of questions must assume that ambiguity will not have a large or systematic effect

on their results" (Fowler, 1992, p. 218). Optimal survey designs can lead to reduced measurement error, higher quality items, and cost savings (e.g., data retrieval methods, administrator training) (Fowler, 1995; Sanchez, 1992). "One of the important realities for students and researchers to grasp is that many of the worst question problems can be identified with simple, informal testing" (Fowler, 1995, p. 153).

A sound research design, coupled with a quality survey, results in data that are accurate representations of the construct(s) of interest. A survey that is an accurate measure produces data that are more useful (Litwin, 2003). Because it is difficult to assess the quality of data collected, it is the survey that is scrutinized for its consistency and accuracy in collecting data (Litwin, 2003). A survey that is well-designed as has "good psychometric properties…can be widely used by other researchers. Therefore, its use can be widespread whether it is well constructed and evaluated" (Coluci, 2012, pp. 122-123). Carmines and Zeller (1979) described a measuring procedure that produced reliable and valid results as "scientifically useful" (p. 6).

Psychometric properties. The quality of a survey has been "defined as the product of the reliability and the validity" (Saris et al., 2010, p. 74). The psychometric properties of a survey enable the researcher to determine the quality of the survey (how good it is), and aid in quantifying otherwise qualitative concepts (Litwin, 2003). "One is commonly advised to take the psychometric quality of the instrument as part of the basis for the choice of a specific test" (Botella, Suero, & Gambara, 2010, p. 11).

The psychometric properties of survey data include evidence of reliability and validity, noted as "twin pillars of psychometric quality" in which validity "holds a preeminent place" (Hogan & Agnello, 2004, p. 802). These two properties go hand-in-hand, often being paired together in textbooks and other publications. Reliable and valid survey data "is vital for both

practitioners and scholars" (Porter, 2011, p. 45); however, these are not "static" properties (Snyder, 2000). Not only is obtaining evidence of the reliability and validity of survey data important (Heiman, 2001), but so is providing evidence of both, as instruments that may be used in future studies need to be verified as both accurate and consistent measures of a specified construct (Coluci, 2012). "If it is reliable and valid, then it has gone a long way toward gaining scientific acceptance" (Carmines & Zeller, 1979, p. 15).

Reliability. There are two types of error that occur in survey research: random error and nonrandom (measurement) error (Carmines & Zeller, 1979; Saris et al., 1998). Random error is that which is unpredictable, "affected primarily by sampling techniques" (Litwin, 2003, p. 5). The less random error there is associated with a survey, the greater the reliability of that survey's data (Carmines & Zeller, 1979). The quality of the data collected from a survey can be enhanced, and the error associated with the survey therefore reduced, if the sample size obtained for the study is increased (Fink, 2003d), or if item stems and response options are appropriately worded and the survey itself is easy to read (Fink, 2003c). Other factors that can reduce the amount of random error can be addressed a priori; including increasing the response rate by oversampling and ensuring the survey for the study is appropriate for the sample (Fink, 2003c, 2003d).

No survey is perfect, resulting in some level of measurement error, which "reflects the precision (or lack of precision) of the survey instrument itself" (Litwin, 2003, p. 6). In other words, survey data that is reliable is that which is free of measurement error (Fink, 2003d; Kline, 2011). Reliability, therefore, is a test of that precision: "the consistency of...measurements when the testing procedure is repeated on a population of individuals or groups" (American Educational Research Association [AERA], American Psychological Association [APA], &

National Council on Measurement in Education [NCME], 1999, p. 25). According to Qualls and Moss (1996), "…measurement…is imprecise. Scores will contain a certain amount of error from one or more sources. It is the presence of that error, quantified by the reliability coefficient, that results in the inconsistency of test scores" (p. 211).

The reliability coefficient indicates the stability and consistency of a survey over time, and is "meaningless unless assessable" (Deming, 1947, p. 147). Because reliability is based on the sample from which it was obtained, evidence of reliability should be calculated for each sample to which the survey is given (Dawis, 1987; Litwin, 2003; Snyder, 2000; Spector, 1992). For those researchers who use an established survey for data collection, providing reliability estimates of the data from the sample of the study is imperative:

Thompson (2003) noted the importance of evaluating score reliability of the data in all studies, "because it is the reliability of the data in hand in a given study that will drive study results, and not the reliability of the scores described in the test manual" (p. 5). Thompson and Vacha-Haase (2000) and Hallinger (2011) concurred in regards to using reliability information from previous studies. The sample of respondents for which reliability was calculated and reported in a manual or previous study is different from the sample used in a dissertation or other study, both most likely obtained from different populations (Kline, 2011; Rodriguez & Maeda, 2006). This is known as "reliability induction (inferring from particular coefficients calculated in other samples to a different population)" (Kline, 2011, p. 69; see also Botella et al., 2010; Green, Chen, Helms, & Henze, 2011; Vacha-Haase, Henson, & Caruso, 2002). More specifically, the characteristics of samples are different; therefore, any estimates of reliability will change with each sample (Kline, 2011; Snyder, 2000; Thompson & Vacha-Haase, 2000). It is advised to "evaluate and report a scale's validity and reliability every time an instrument is

administered" (Barry et al., 2013, p. 2), "even if the focus of [the] research is not psychometric" (Wilkinson & the APA Task Force on Statistical Inference, 1999, p. 596).

Evidence of reliability "should be a routine part of any survey project" (Fowler, 1995, p. 148). It can be established using multiple methods that are based on classical test theory, which is used to assess random measurement error (Carmines & Zeller, 1979; Netemeyer et al., 2003; Spector, 1992): test-retest, alternate (alternative) forms, and internal consistency. These methods utilize a correlation analysis to assess the extent of the relationship between the survey's items to each other or to a comparable survey's items. Although the test-retest method is commonly used (Litwin, 2003), it is not recommended for survey research due to being prone to practice effects (Carmines & Zeller, 1979; Litwin, 2003; Netemeyer et al., 2003; Salkind, 2006). The alternate forms method is highly recommended as it eliminates the practice effect that can occur in test-retest, given that the construction of the forms is carried out well (Carmines & Zeller, 1979; Litwin, 2003). However, the split-halves version of alternate forms reliability is also a method that is not recommended, as the degree of reliability can change depending on how the items were split into halves (Carmines & Zeller, 1979). The internal consistency method is more popular in survey research and is more often used (Botella et al., 2010; Netemeyer et al., 2003), as it is simple, cost- and time-effective, and can be a quick assessment of reliability estimates.

Internal consistency. The internal consistency method of reliability is used to intercorrelate all items on an instrument administered one time to one sample; this method is not used for a single item, but for a single survey or a group of items used to measure a single construct (Carmines & Zeller, 1979; Gliem & Gliem, 2003; Kitchenham & Pfleeger, 2002; Litwin, 2003; Netemeyer et al., 2003; Spector, 1992). The internal consistency coefficient (e.g., Cronbach's coefficient alpha) indicates how consistently the items measure the same variable or

construct within the measure (i.e., homogeneity of the items) (Benson & Clark, 1982; Carmines & Zeller, 1979; Coluci, 2012; Kline, 2011; Litwin, 2003; Netemeyer et al., 2003; Salkind, 2006; Spector, 1992). This highly recommended technique (Carmines & Zeller, 1979) is most likely used in survey research moreso than test-retest and alternate forms methods due to its simplicity and ease of interpretation.

The value of a coefficient ranges between 0 and 1 (Gliem & Gliem, 2003; Kline, 2011; Newton & Rudestam, 2013; Spector, 1992). A high coefficient indicates strong internal consistency of the items; a low value can be improved by adding more items to the survey (higher alpha values are associated with increasing the number of items [Netemeyer et al., 2003]), deleting items that do not correlate with others, or the investigator can reexamine and further clarify the existing items (Litwin, 2003; Spector). Verifying the internal consistency of survey items can be on ongoing process (Spector, 1992).

A coefficient value of at least .80 is acceptable evidence of internal consistency (Benson & Clark, 1982; Gliem & Gliem, 2003), and is a reasonable goal for newly-developed surveys (Netemeyer et al., 2003). Regarding the internal consistency of items, the following can be used to interpret a coefficient: around .90 excellent; .80 very good or good; .70 acceptable or adequate; and less than .50 unacceptable, as these values indicate the amount of precision in the research is not adequate due to too much random error (Gliem & Gliem, 2003; Kline, 2011). Gliem and Gliem (2003) further recommended coefficients around .60 to be interpreted as questionable and those around .50 as poor when using a rating scale. Newton and Rudestam (2013) warned "that the context of the study is crucial in determining the relative importance of the size of the correlation coefficient" (p. 299). Correlation coefficients can be interpreted in

terms of strength, in which coefficients in the vicinity of .80 are strong, in the area of .50 are moderate, and in the proximity of .20 are weak (Newton & Rudestam, 2013).

"One of the major drawbacks of new survey instruments is that they are often nothing more than collections of questions that seem to the surveyors to fit well together" (Litwin, 2003, p. 25). Poor reliability can compromise "the ability of a study to yield noteworthy effects" (Thompson, 2003, p. 5). Poor score reliability weakens effect-size magnitudes, and that only reporting reliability coefficients from previous studies is insufficient (Kieffer et al., 2001; Snyder, 2000). "When one modifies an instrument or combines instruments in a study, the original validity and reliability may not hold for the new instrument, and it becomes important to reestablish validity and reliability during data analysis" (Creswell, 2014, p. 160). Unfortunately, this is not completed because researchers may not know or realize that the reliability of their own scores affects validity, power of statistical tests, and effect sizes (Botella et al., 2010; Kieffer et al., 2001; Kline, 2011). It is essential for an investigator to realize that measurement error can impact effect sizes (Snyder, 2000; Thompson, 1994a; Thompson & Vacha-Haase, 2000) and validity (Fowler, 1995), which "depends on the extent of nonrandom error present in the measurement process" (Carmines & Zeller, 1979, p. 15). Reliability is important in order to gain information about the validity of the scores, and although it is easy to collect information about the consistency of scores, "evaluating the validity of questions is not easy" (Fowler, 1995, p. 148).

Validity. Validity is "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9), more commonly known as evidence that a survey "measures what it purports to measure" (Carmines & Zeller, 1979, p. 12; Fink, 2003d, p. 50). Validating a survey is an ongoing process that involves

specifying the constructs and their theoretical relationships, examining the relationships between the measures (tests, instruments, surveys) of these constructs, and interpreting the evidence of the relationship between the measures (Benson & Clark, 1982; Carmines & Zeller, 1979; Coluci, 2012; Netemeyer et al., 2003; Spector, 1992). A survey can be validated if it "relates to other measures consistent with theoretically derived hypotheses concerning the concepts (or constructs) that are being measured" (Carmines & Zeller, 1979, p. 23). An investigator who creates their own survey should validate it to ensure the survey is compatible with an established survey of sound quality (Alvarez et al., 2012). A positive relationship is evidence of validity; however, a negative relationship would require further investigation. A negative relationship between two instruments could mean that the instrument in question lacks validity of the construct *of interest*; it could be a measure of another construct (Carmines & Zeller, 1979). Other interpretations of a negative relationship between two instruments could imply the relationships between constructs is incorrect, the method used to correlate the constructs is incorrect, or there is overall a lack of validity or reliability from the data (Carmines & Zeller, 1979).

If an established survey does not exist for comparison, a committee of subject matter experts is employed to review the survey and its items (Benson & Clark, 1982; Coluci, 2012; Dawis, 1987; Litwin, 2003). The extent to which the experts agree on the clarity, inclusion, number, adequacy, and relevancy of survey items is considered evidence of validity (Coluci, 2012). Coluci (2012) recommended a 90% agreement rate of the experts.

Another method of obtaining evidence of validity is through exploratory or confirmatory factor analysis, in which interrelated variables are identified in a set of data (Benson & Clark, 1982; Carmines & Zeller, 1979; Coluci, 2012; Snyder, 2000; Spector, 1992). In the event that a

survey is constructed to measure more than one concept, factor analysis is recommended as a method of identifying those multiple concepts, as it does not assume the items are parallel measures (Carmines & Zeller, 1979; Spector, 1992). This method can also be used in the survey development process, in which factor analysis aids the investigator in choosing items to be included in the survey that measure the construct(s) of interest (Dawis, 1987; Litwin, 2003; Spector, 1992).

Not only is validating an instrument not easy (Fowler, 1995; Kitchenham & Pfleeger, 2002), its importance cannot be stressed enough. "No one can be found to say that test validity is not important" (Shultz, Riggs, & Kottke, 1998, p. 266). Fowler (1995) and Thompson (2003), among others, have emphasized the importance of validation. Fowler (1995) stated there is a "need to continue to encourage researchers routinely to evaluate the validity of their measurement procedures from a variety of perspectives... [and there is a] need to develop clear standards for what validation means for particular analytic purposes" (p. 154). Thompson (2003) supported this, stating that "questions of validity are important in research studies, because our inferences regarding study outcomes will be compromised if our scores are invalid" (p. 6). The evaluation of validity is only as good as the items used to validate the instrument (Fowler, 1995). According to Netemeyer et al. (2003), validity "is not assessed directly but is inferred from…the quality of the procedures that were employed in the development and validation of the measure" (p. 71).

Survey data that is not valid can lead to incorrect inferences and misleading conclusions and recommendations regarding the survey measuring the construct(s) of interest (Barry et al., 2013; Carmines & Zeller, 1979). Unfortunately, in survey research of subjective constructs, "many of the most interesting and important things we ask people to report are virtually

impossible to validate" (Fowler, 1995, p. 142). According to Carmines and Zeller (1979) and Spector (1992), the use of validation procedures is limited in the social sciences when the content included in a survey addresses abstract concepts that are difficult to define. Subjective constructs, such as those used in the social sciences and other fields, are difficult to evaluate as there is no comparison to be made other than to a respondent's self-report (Fowler, 1995; Spector, 1992). For theoretical constructs that are difficult to evaluate, "evidence is collected to either support or refute validity" (Spector, 1992, p. 46). The validity of data from a subjective measure can be improved by ensuring the items are reliable, which transpires through improved item design (Fowler, 2009).

Perfect reliability and validity is not achievable, as measurement of these properties is the extent of the degree that each is present (Carmines & Zeller, 1979; Litwin, 2003). Both properties are functions of scores, not functions of the survey itself (Barry et al., 2013; Benson & Clark, 1982; Hogan & Agnello, 2004; Kieffer et al., 2001; Kline, 2011; Snyder, 2000; Thompson, 1994a, 1994b, 1994c, 2003; Wilkinson & the APA Task Force on Statistical Inference, 1999), and many have stressed that authors move away from stating a test is reliable, to the test *scores* are reliable (Green et al., 2011; Hogan, Benjamin, & Brezinski, 2000; Thompson & Snyder, 1998; Thompson & Vacha-Haase, 2000; Vacha-Haase et al., 2002; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). The reliability of scores is necessary for validity, but it is not a sufficient condition (Kline, 2011; Netemeyer et al., 2003; Qualls & Moss, 1996; Salkind, 2006; Thompson, 2003). Reliability does, however, affect the validity of scores in that if the scores are inconsistent, it is implied the scores are also not valid; however, the opposite is not necessarily true (Fowler, 1995; Thompson, 2003). If data are not consistently measured by a survey, the survey may not be an accurate measure of the content of interest and therefore not a

predictor of future performance. According to Qualls and Moss (1996), "validity evidence and reliability evidence are by far the two most crucial elements that underlie judgments regarding the quality of scores derived from instruments" (p. 211). However, there is extensive research on the lack of reporting these properties:

> there is substantial literature to show that the methods used in conducting survey research can significantly affect the reliability, validity, and generalisability of study results. Without clear reporting of the methods used in surveys, it is difficult or impossible to assess these characteristics. (Bennett et al., 2011, p. 2)

Providing evidence of these elements is essential, and "recognized as a professional responsibility" (Shultz et al., 1998, p. 266). A researcher who creates a survey for data collection, or modifies an existing survey, should fully report evidence of validity and reliability (Draugalis et al., 2008; Fowler, 2009; Rodriguez & Maeda, 2006).

Lack of psychometric reporting. Studies have shown that together or individually, evidence of reliability and validity is not being reported in dissertations or published research, no matter the type of research design or methodology used. Thompson (2003) noted that "given the diversity of participants across studies, simple logic would dictate that authors of *every* study should provide reliability coefficients on the scores for the data being analyzed, even in nonmeasurement [sic] substantive inquiries" (p. 9). The same is implied for providing evidence of validity: "without validity, all the other measurement characteristics become relatively inconsequential" (Hogan & Agnello, 2004, p. 802). Regardless of the type of instrumentation used for data collection, evidence of validity and reliability should be reported so it is known the instrument is consistently measuring what it purports to measure. Unfortunately, there is an abundance of research on the lack of psychometric reporting.

In his review of dissertations, Thompson (1994a) found that students were continuing with data analysis even though the results of reliability analyses were questionable (implying questionable validity); argued for needed reliability analyses, but did not conduct it themselves; and used bad language when referring to reliability. Most troubling, Thompson (1994a) noted, was the "pattern where students do not analyze the reliability of their scores, when their scores are actually not very reliable, and these problems are not considered during analysis and/or interpretation" (p. 7). Thompson (1994c) stated that "we should expect authors of published research to offer empirical evidence that the scores they are actually analyzing have reasonable measurement integrity" (p. 1), previously remarking that "it is axiomatic that measurement integrity is vital in quantitative research" (Thompson, 1988, p. 33). Karadag's (2011) analysis of dissertations revealed insufficient quality of psychometric reporting, due to multiple mistakes including not providing adequate information about results and erroneously deleting items after analyses.

Such reporting, or lack thereof, of psychometric evidence is seen in published literature, in which reliability, validity, or both were not reported for the instrumentation used (Barry et al., 2013; Bennett et al., 2011; Draugalis et al., 2008; Goodwin & Goodwin, 1985a, 1985b; Hogan & Agnello, 2004; Hogan et al., 2000; Kieffer et al., 2001; Meier & Davis, 1990; Qualls & Moss, 1996; Thompson & Snyder, 1998; Tinsley & Irelan, 1989; Vacha-Haase et al., 1999; Willson, 1980). Meier and Davis (1990) stated the "failure to provide estimates and thereby demonstrate the adequacy of scales' psychometric properties leaves research results open to an alternative explanation, that is, the scales used are improper measures of the research constructs" (p. 114). Providing evidence of validity and reliability is the responsibility of the investigator who developed the survey (Qualls & Moss, 1996). "Studies missing this information have diminished

credibility and usefulness and generally provide poor examples of the research process to graduate students and other researchers" (Goodwin & Goodwin, 1985a, p. 18).

The Affect of Survey Quality on Results

Each aspect of a survey (research design, objectives, items, pilot test methods, psychometric properties, and methodology) impacts others and the survey's "precision, accuracy, and credibility" (Fowler, 2009, p. 1). The quality and components of a survey affects the results, conclusions, and interpretations made in a study. "Whether the design is flawed determines whether the data is flawed, which determines whether the conclusions of the study are flawed" (Heiman, 2001, p. 20). The manner in which an item is written, the response options offered, the number of items on the survey, and the instructions given to participants all influence the data that will be obtained (Kitchenham & Pfleeger, 2002). "At a minimum, valid survey findings depend on clearly stated purposes, justified samples, accurate data collection, and appropriate statistical analysis and interpretation" (Fink, 2003a, p. 34). Fowler (1995) discussed the connection between data, validity, and results reporting in survey research:

> [it is] important that researchers continue to evaluate the quality of reporting resulting from surveys whenever possible...the results from such validating efforts provide critical evidence of the quality of survey data. They provide a stimulus to researchers to continue to work on improved measurement. They serve as important reminders to users of survey data about the appropriate uses and the limits to which survey measurement can be put. (p. 148)

The continuation of evaluating a survey and its components results in quality data and findings: "the quality of data will be no better than the most error-prone feature of the survey design"

(Fowler, 2009, p. 7) and "the results of the study will only be as good as the instrument" (Lunenburg & Irby, 2008, p. 32).

Calls for action have been made to evaluate and remedy problems within all aspects of survey research (Fincham & Draugalis, 2013; Lehman, 1974). The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) state that instruments used only for research purposes are not subject to the "standard test development functions that occur prior to the operational use of a test" (p. 48). However, should studies that employ the use of statistical analyses be subjected to the rigors of such standards?

Particular facets of survey research are crucial to report, namely psychometric properties, so readers of the studies can accurately interpret the results and critically examine the interpretations made by the authors. Further, providing evidence of validity and reliability impacts the use of a survey in future studies. There are many producers and consumers of survey research who lack the background, expertise, and scientific attitude necessary to conduct competent research or the perspective to interpret its results, which can lead to misrepresentations or fabrications (Fowler, 1995; Miller, 2010). "Researchers now may be entirely untrained individuals who use cheap software to construct 'do it yourself' surveys, relying on volunteer respondents" (Miller, 2010, p. 604). Each component of survey development is crucial to, and affects the quality of, the survey, data collected, results, and interpretation of the results. "Bad surveys produce bad data – that is, data that are unreliable, irreproducible, or invalid, or that waste resources. Good surveys, on the other hand, yield critical information and provide important windows into the heart of the topic of interest" (Litwin, 2003, p. 1).

Just as published literature is used to advance fields of study, surveys can be used to advance all facets of survey research. According to Bennett et al. (2011), "poor reporting compromises both transparency and reproducibility, which are fundamental tenets of research" (p. 9). Setting requirements of psychometric reporting, and increasing expectations of quality survey research, should commence in graduate school. If it has been shown that faculty members do not expect graduate students to make a significant contribution to their field of study (Lovitts, 2007), is a similar effort put towards the development, implementation, and evaluation of a student-developed survey for a doctoral dissertation? A study of the quality of the surveys used in doctoral dissertations is apparent, as catching errors early in careers may decrease the number of studies conducted on the quality of research in refereed journals. "If research findings are to be used in determining educational practice and further research, some attempt to analyze and evaluate the various aspects of a report are vital" (Kohr & Suydam, 1970, p. 78).

Summary

Studies of the quality of research have focused on dissertations and refereed journal articles, specifically regarding psychometric reporting. The increased use of surveys in dissertation research, as well as the increase in student-developed instruments, justifies a review of the surveys used for data collection. Chapter three includes methodological details of the current study.

CHAPTER THREE

METHODOLOGY

The purpose of this study was to examine student-developed survey items and rating scales in doctoral dissertations, specifically to address survey quality and psychometric reporting practices. In this chapter are details on the research design of the study, including a description of the sample, sampling procedure, literature search strategies and criteria for inclusion in the study, and the method of data collection. The data analysis section is a thorough description of the methods and guidelines used to review the surveys and relevant components. Assumptions and limitations of the study are addressed throughout these sections.

Research Design

This study was exploratory and descriptive in nature. Student-developed surveys from doctoral dissertations were sampled and chosen for inclusion as data for the study using set guidelines, as detailed in the data collection section. The data were subjected to thorough review to encompass how doctoral students created, implemented, and reported essential components of their surveys with rating scales.

Sample. The population of interest included all doctoral dissertations that included at least one student-developed survey. Due to the increased use of survey research in dissertations, and the frequency of those surveys being developed by the researchers, focus was placed on surveys that were created by doctoral students specifically for their dissertation research.

Sampling procedure. Dissertations including surveys that were self-administered and used a direct collection method were reviewed, as this type of data collection was the focus of the current study. This excluded dissertations that used interview, observation, and record review methods of data collection. A sample was collected from the population of dissertations using broad, albeit strict, sampling guidelines. A dissertation was included in the sample if it

satisfied the following criteria: 1) it was a published or public dissertation; 2) it employed the use of at least one student-developed survey; 3) the survey was included with the dissertation (as recommended by Creswell, 2014; Fowler, 2009); 4) the dissertation was published in the United States; and 5) the dissertation was published between 2009 and 2013, to include the most current data on students' practices. Limiting the sample to specific fields of study were not included in the sampling criteria. It was assumed the sample of dissertations was representative of the population of all dissertations that met the specified criteria.

Literature search strategy. To locate the doctoral dissertations for the sample, a database search strategy was employed. The ProQuest Dissertations and Theses (PQDT) database was searched, which is a collection of theses and dissertations from universities around the world. Using an *Advanced Search*, the following search criteria were used:

- Limit to: Full-text
- Publication date: *Specific date range*
    - Start year: 2009
    - End year: 2013
- Manuscript type: Doctoral dissertations
- Language: English

Keyword terms used to search PQDT included

- "researcher-developed survey" OR "investigator-developed survey" OR "self-developed survey"
- AND "Likert" OR "rating scale"

(both with and without hyphens) *anywhere* in the dissertation text were pulled from the PQDT database. This resulted in 402 dissertations. A second preliminary search using

- "researcher-designed survey" OR "investigator-designed survey" OR "self-designed survey"
- AND "Likert" OR "rating scale"

(both with and without hyphens) *anywhere* in the text resulted in an additional 89 dissertations, after PQDT removed duplications between the two searches. The use of "student" as a keyword did not produce additional results.

The potential sample of 491 dissertations was then reviewed using additional criteria. Each dissertation was searched to verify the author used a self-developed survey for data collection, the survey created was not one that solely consisted of items to collect demographic information, the survey was included with the dissertation, and the survey included a rating scale. A further criterion was that the survey's rating scale had to include at least three response options (i.e., the rating scale was not dichotomous). Search terms used to locate the dissertations were found in the literature reviews or other areas of the dissertations, not necessarily reflective of the type of data collection tool used in the study. These dissertations were eliminated from the potential sample.

After the secondary review, 93 dissertations were eliminated from the sample. These dissertations were removed because the author solely used existing, or modified existing, survey items ($n = 44$); others were removed because it was unclear in the writing which survey items were student-developed ($n = 21$). Additional dissertations were eliminated because they did not fit the specified criteria: the survey items were not on a rating scale with three or more options ($n = 9$); the survey was not developed for the dissertation ($n = 7$); the full dissertation was not available ($n = 1$); and the dissertation was not published in the United States ($n = 1$). Nine dissertations were removed because the survey was illegible ($n = 2$); the instrument was not

applicable (n = 6), such as it was an evaluation form, vignette, or was scenario-based; or only one item was used throughout the survey (n = 1). One author, as well as her doctoral program and committee, was known; therefore, her dissertation was eliminated from the data so as not to introduce any possible bias in the review of her survey and other dissertation components. Eliminating irrelevant dissertations resulted in 246 dissertations to be included as the sample for the current study

Data collection. The following data were collected to generally describe the sample of authors and their doctoral dissertations: year of publication on PQDT, highest degree obtained prior to the terminal degree, type of terminal degree obtained (e.g., Ph.D.), field of study, educational institution, and research level of that institution. In addition to reviewing the student-developed surveys using the specified item writing and rating scale guidelines, the number of surveys in each dissertation was recorded, as well as the total number of items in the survey and the number of survey items included in analyses of the items. In the case a component of a survey's items or scale was labeled as not including information or not satisfying a guideline, the reasoning behind this label was recorded for additional descriptive information.

Identified in each dissertation was evidence of whether the author included some sort of blueprint in the dissertation, such as a matrix of survey items and research questions, or specification of the survey's purpose, objectives, or constructs measured. The content of the dissertations was also reviewed to evaluate whether a pilot test of the survey items was conducted, and whether reliability and validity analyses were performed. Each dissertation was reviewed for the method of obtaining reliability evidence, whether the correlation coefficient for reliability was reported, and the value of the coefficient. The method of obtaining validity evidence and whether evidence for validity was reported was also collected. Validity evidence

was considered when reviewing descriptions of the survey development process, in which aligning survey objectives to literature review or conducting a pilot test of the survey prior to data collection were considered methods of validation. Included in Appendix A is a list of sources that emphasized the need for inclusion of a survey blueprint, pilot test, and psychometric evidence.

Data analysis. Because this study is exploratory, all data were aggregated and reported descriptively (see Table 1). The characteristics of the doctoral authors and their dissertations were counted and reported using frequencies or percentages of occurrence. The research levels of the academic institutions were labeled using the *Basic Classification* categories of doctorate-granting universities according to The Carnegie Classification of Institutions of Higher Education™ (http://carnegieclassifications.iu.edu/). Included in Appendix B is the schema used for coding the items and rating scales.

Table 1

Results Reporting Methods

Component	Reporting Method
Year of publication	Years and frequencies of each.
Highest degree obtained	Types and frequencies of each.
Terminal degree obtained	Types and frequencies of each.
Field of study	Types and frequencies of each.
Institutional research level	Types and frequencies of each.
Survey blueprint	Percentage identifying alignment between survey and study.
Items	Average percentage of items that broke item writing guidelines.
Rating scale(s)	Frequency of occurrence for each guideline.
Pilot test	Percentage that completed a pilot test.
Technique	Methods of pilot-testing and frequencies of each.
Reliability	Percentage that conducted reliability analysis.
Types of analysis	Types and frequencies of each.
Results reported	Frequencies of coefficients in each range.
Validity	Percentage that conducted validity analysis.
Types of analysis	Types and frequencies of each.

Note. See Appendix A for references.

The survey blueprint category was reported as the percentage of dissertations in which the objectives of the survey were identified, or the alignment between the survey items and research questions of the study was described. Within the narrative of the dissertations, evidence of pilot-testing the survey items was explored. The methods of pilot-testing the survey items (i.e., using a representative sample) were reported with corresponding frequencies for each method.

The percentages of dissertations in which one or both psychometric properties were reported were calculated. The types of reliability and validity analyses were reported, with corresponding frequencies that occurred within each (i.e., only reliability, only validity, both

reliability and validity, neither reliability nor validity). The frequency of authors who reported a reliability coefficient of at least .80 was recorded (Benson & Clark, 1982; Gliem & Gliem, 2003; Netemeyer et al., 2003), as well as the frequencies of coefficients in each of the following ranges:

- Strong: ≥ .80
- Moderately strong: .60 to .79
- Moderate: .40 to .59
- Moderately weak: .20 to .39
- Weak: < .20

The surveys were subjected to a thorough review based on the characteristics of quality survey items and rating scales presented in chapter two from a review of relevant literature. Each dissertation survey was reviewed using the *Guidelines for Reviewing Survey Items and Rating Scales* (see Table 2), in which the surveys' items and rating scales were assessed. Although more characteristics were presented in chapter two than what is presented in the table, only those characteristics that could be objectively evaluated are included in Table 2.

Table 2

Guidelines for Reviewing Survey Items and Rating Scales

Component	Guideline	Description	References
Stem	Accurate mechanics	Correct syntax, punctuation, and other mechanics according to APA (2010).	Fink, 2003d; Fowler, 2009; Ho, 2005
	Appropriate length	Introduction of a definition or description of a concept occurs prior to the question or statement posed.	Benson & Clark, 1982; Converse & Presser, 1986; Fink, 2003d; Fowler, 1995
	Double-barreled	Question or statement about a single concept (is mutually exclusive or does not include multiple concepts).	Converse & Presser, 1986; Fowler, 1995, 2009; Heiman, 2001; Likert, 1932; Netemeyer et al., 2003; Spector, 1992
	Positively worded	Exclusion of negative words or terms (i.e., *except*, *not*).	Converse & Presser, 1986; Fink, 2003d; Spector, 1992
Rating Scale	Scale starts with 0 or 1	The rating scale begins with a 0 or 1 (if values are used).	Schaeffer & Presser, 2003
	Continuous scale	The scale and its response options are along a continuum; options are not out of order or missing.	Fink, 2003d; Fowler, 1995, 2009; Spector, 1992
	One- or two-sided scale	A one-sided scale is used (positive [zero to positive values] or negative [negative values to zero]), or a two-sided scale is used (negative to positive values).	Dawis, 1987; Fink, 2003d; Fowler, 1995; Spector, 1992
	Side of scale presented first	Positive or negative side of two-sided scale presented first.	Fowler, 1995; Spector, 1992
	Symmetrical	Options on both sides of the midpoint of the two-sided scale are symmetrical (i.e., match across the scale).	Fink, 2003d; Fowler, 1995; Spector, 1992
	Number of options	Number of options provided in the scale.	Fowler, 1995; Netemeyer et al., 2003; Saris et al., 2010; Spector, 1992
	Type of options	Anchor, value, or both presented.	
	Anchors are appropriate	Anchor wording matches stem wording; the anchors are continuous across the scale (i.e., same adjectives).	Netemeyer et al., 2003; Saris & Gallhofer, 2007
	Each scale point is labeled	Each scale point is labeled with an anchor or value. If both are used, they are not contradictory.	Krosnick, 1999; Netemeyer et al., 2003; Saris & Gallhofer, 2007; Schaeffer & Presser, 2003
	Neutral option	Inclusion of a neutral response option (e.g., *Don't Know*, *Neither Agree nor Disagree, Neutral, No Opinion*).	Converse & Presser, 1986; Dawis, 1987; Fink, 2003d; Fowler, 1995, 2009; Heiman, 2001; Krosnick, 1999; Likert, 1932; Saris & Gallhofer, 2007; Spector, 1992

This tool was evaluated by subject matter experts for clarity, appropriateness, ease of use, and validity. In the event a doctoral student used more than one self-developed survey in the dissertation, the multiple surveys' item characteristics were reported aggregately and rating scale characteristics were reported separately.

Each item from the surveys were subjected to the four item-writing guidelines in Table 2. The percentage of items within a survey that did not adhere to one or more of these guidelines was calculated for each dissertation. The total number of items in the survey was multiplied by the number of item writing guidelines. For example, a survey with 10 items had a total possible score of 40 (e.g., 10 survey items multiplied by four item-writing guidelines). The total number of guidelines that was not adhered to by one or more items in the survey was then divided by the total possible score (i.e., 40). This percentage was calculated for each dissertation survey; the percentages were reported as an average for all surveys. The percentages of items that were recorded for each item writing guideline were reported.

Each survey's rating scale was subjected to the rating scale guidelines in Table 2. Rating scales that included at least three response options were included in analysis. Because the rating scale guidelines are not necessarily right or wrong approaches, the dissertation surveys were reported aggregately according to their membership in the categories for each guideline (see Table 2).

Reliability. Due to the potential of introducing subjective review in data analysis, a random sample of 10% of the original 246 dissertations was drawn for a second analysis (n = 25) approximately two weeks after the initial analysis was completed. To ensure consistency in coding of the dissertations, the random sample was evaluated and compared to their original

categorizations and scores. It was deemed after the second review that components of the dissertations were consistently coded, as there was 100% agreement between analyses.

Summary

The methodology of this study encompassed inclusion criteria of doctoral dissertations with student-developed surveys. Content within the dissertations, as well as guidelines for reviewing the survey items and rating scales, was detailed. Presented in chapter four are the results of data analyses.

CHAPTER FOUR

RESULTS

A sample of 246 doctoral dissertations was included in data analysis. Included in Table 3 is the number of dissertations published between 2009 and 2013. The publication years were those reported by PQDT, not necessarily the year of the author's defense.

Table 3

Dissertation Sample by Year

Year	*n*	%
2009	50	20.3
2010	48	19.5
2011	46	18.7
2012	52	21.1
2013	50	20.3
Total	246	100.0

The research levels of authors' institutions were predominantly Doctoral/Research Universities and Research Universities with high research activity. Frequencies of the research levels are presented in Table 4.

Table 4

Research Levels of Authors' Institutions

Research Level	*n*	%
RU/VH: Research Universities (very high research activity)	39	15.90
RU/H: Research Universities (high research activity)	77	31.13
DRU: Doctoral/Research Universities	80	32.50
Master's L: Master's Colleges and Universities (larger programs)	42	17.10
Master's S: Master's Colleges and Universities (smaller programs)	1	.40
Spec/Faith: Theological seminaries, Bible college, and other faith-related institutions	4	1.60
Spec/Med: Medical schools and medical centers	1	.40
Not found	2	.80

If identified, highest degrees obtained prior to the terminal degree was predominantly a master's degree (n = 86). Seven authors held an Educational Specialist (Ed.S.) degree. Approximately half (n = 121) of the authors received a Doctor of Education (Ed.D.) degree upon completion of the dissertation. Forty-one percent (n = 101) received a Doctor of Philosophy (Ph.D.) upon completion. Seven authors received a Doctor of Business Administration (D.BA.), five received a Doctor of Ministry (D.Min.), three received a Doctor of Management (D.Mgt.), and two received a Doctor of Musical Arts (DMA). Programs in psychology (Psy.D.), social work (DSW), nursing science (DNS), public administration (DPA), musical arts (DMA), human environment (DHE), and health administration (DHA) each had one. One degree was not specified in the dissertation.

The majority (n = 190) of authors' fields of study was education; 21 were business degrees. Other fields of study included religion (n = 7), psychology (n = 6), music (n = 5),

nursing (n = 3), public policy and administration (n = 4), and leadership and technology (n = 2). Communications, forest resources, health sciences, leadership and foundations, library science, public health, social research, and social work each had one.

Survey Items and Rating Scales

There were 278 surveys created by the authors for the 246 dissertations. Included in data analysis were 262 surveys, as 16 authors included the same surveys for different sample groups. The majority of authors created one survey (n = 220), 21 created two surveys, four created three surveys, and one created four surveys.

Included in the 262 surveys were 7,027 items which were evaluated using the specified item writing guidelines. The range of the total number of items in the surveys (6 to 203) differed from the number of survey items included in data analysis (2 to 177). Reasons for the discrepancy were due to the surveys including demographic items, modified items, items not on a rating scale, or some combination of these. Regardless of the number of items identified by the author within the dissertation, items that were grouped within a survey were assessed individually (i.e., one item stem comprised of multiple phrases).

All but 13 of the surveys included items that broke at least one item-writing guideline. The average percentage of items that broke item writing guidelines was 88.93% (SD = 7.475), with a range of 63.33% to 100%. The results of the survey item analysis is presented in Table 5. (See Appendix B for the schema used to code item writing properties.)

Table 5

*Items that Followed Survey Item Guidelines (*n = *7,027)*

Guideline	*n*	%
Accurate mechanics	5,995	85.31
Appropriate length	6,919	98.46
Not double-barreled	5,240	74.57
Positively worded	6,863	97.67

Within the 262 surveys, 480 rating scales were evaluated using the specified rating scale guidelines. The number of scales in a survey was characterized not only by differing numbers of response options, but by differing types of response options. For example, a dissertation was coded as including two rating scales if the survey included two types of scales (i.e., one agreement scale and one frequency scale), or one type of scale with two sets of response options (i.e., one agreement scale with three and five response options). The majority of surveys (*n* = 153) included one scale, 43 included two scales, 21 included three scales, and 13 included four scales. Others included five (*n* = 4), six (*n* = 5), seven (*n* = 2), nine (*n* = 1), 10 (*n* = 1), and 13 (*n* = 1) scales. In two surveys, 15 types of scales were included.

The agreement scale was the most commonly used (*n* = 123): 104 authors used it for the entirety of their surveys and nine used it in combination with another anchor. Frequency scales were used in 33 surveys; 21 used it for the entirety of the survey and two used it in combination with another anchor. Ten authors used a combination of agreement and frequency scales. Thirty surveys included some other types of response anchors worded specifically to the items posed (i.e., degree of confidence). Sixty-nine authors used various combinations of response options within their surveys.

The number of response options in surveys that included only one rating scale throughout the entire survey (n = 195) are shown in Table 6.

Table 6

*Number of Response Options in Surveys with Use of Only One Rating Scale (*n = *195)*

Number of Options	*n*	%
3	5	2.56
4	31	15.90
5	116	59.49
6	27	13.85
7	12	6.15
10	4	2.05

Using combinations of three to 13 options, 51 authors used up to four numbers of response options in their rating scales. Table 7 includes the frequencies of each occurrence in all surveys (n = 310), which is inclusive of the data in Table 6.

Table 7

*Number of Response Options in All Rating Scales (*n = *310)*

Number of Options	*n*	%
3	20	6.45
4	61	19.68
5	148	47.74
6	50	16.13
7	18	5.81
8	2	.65
9	3	.97
10	7	2.26
13	1	.32

Results of the rating scale analysis for each dissertation (n = 246) are presented in Table 8.

Table 8

*Rating Scale Analysis Results for Each Dissertation (*n = *246)*

Guideline	Category	*n*	%
Scale starts with 0 or 1	Starts with 0	5	2.03
	Starts with 1	108	43.90
	Both used	1	.41
	No values used (N/A)	130	52.85
	Both N/A and started with 1	2	.81
Continuous scale	No	10	4.07
	Yes	230	93.49
	Both	6	2.44
One- or two-sided scale	One-sided negative	0	-
	One-sided positive	74	30.08
	Two-sided	119	48.37
	Both one-sided positive and two-sided	53	21.55
Type of options	Anchor	126	51.22
	Value	1	.41
	Both	109	44.30
	Anchor and both	10	4.07
Anchors are appropriate	Not appropriate	22	8.94
	Appropriate	217	88.21
	Both	7	2.85
Each scale point is labeled	No	25	10.16
	Yes	216	87.81
	Both	5	2.03
Neutral option	No	111	45.12
	Yes	102	41.46
	Both	33	13.42

Scales were considered not continuous when options were out of order or missing. For those 30 that did not have each scale point labeled: nine were labeled at mid- and end-points, 20 were labeled only at the end-points, and the neutral option for one scale was not identified. Anchors were deemed inappropriate if the scale did not match the item posed (i.e., a frequency scale was given for an agreement item), if various response options (i.e., adjectives) were used across the scale, or if there was no discernible difference between options.

Properties relevant to those surveys that included a two-sided scale (n = 172) are presented in Table 9.

Table 9

*Two-Sided Rating Scale Properties (*n = *172)*

Guideline	Category	*n*	%
Side of scale presented first	Negative	95	55.23
	Positive	73	42.44
	Both used	4	2.33
Symmetrical	Not symmetrical	8	4.65
	Symmetrical	162	94.19
	Both identified	2	1.16

Response options were deemed not symmetrical if response options did not match on both sides of the scale. For example, one scale was labeled as *Strongly Agree*, *Agree*, *Somewhat Disagree*, and *Disagree*. (See Appendix B for the schema used to code rating scale properties.)

Survey Development and Psychometric Properties

The majority (n = 177) of authors identified the alignment of the survey to the purpose of the study, through a description of a survey blueprint or a matrix of tying survey items to research questions. The majority (n = 171) pilot-tested the surveys prior to administration, of which most used a representative sample. The types of pilot test methods and the frequency of

each are reported in Table 10. (See Appendix C for the schema used to code methods of pilot testing, reliability, and validity.)

Table 10

*Frequencies of Pilot Test Methods (*n *= 171)*

Technique	*n*	%
Sample only	135	78.95
SME review only	25	14.62
Both	1	.58
Not identified	10	5.85

Note. SME = Subject Matter Expert

Those coded as not identified were authors who mentioned a pilot test in the methodology, but did not provide detail as to how it was performed.

Validation of the surveys was coded as being completed if the author stated validity in some capacity beyond discussion of a pilot test, if applicable. Seventeen authors did not conduct a pilot test, report evidence of reliability, or report any method of validity; 75 reported all three (see Table 11). Excluding the use of a pilot test, 52 authors did not present evidence of reliability or validity.

Table 11

Matrix of Psychometric Reporting

	No Pilot Test		Pilot Test		Total	
Reported	*n*	%	*n*	%	*n*	%
Only validity	36	14.63	44	17.89	80	32.52
Only reliability	2	.81	17	6.91	19	7.72
Both	20	8.13	75	30.49	95	38.62
Neither	17	6.91	35	14.23	52	21.14
Total	75	30.49	171	69.51	246	100.00

Table 12 represents a matrix of psychometric reporting by the research levels of authors' institutions.

Table 12

Matrix of Psychometric Reporting by Research Level of Institutions

Research Level		Only Validity	Only Reliability	Both	Neither	Total	Overall Total	
							n	%
RU/VH	No pilot test	4	1	3	3	11	39	15.85
	Pilot test	5	5	10	8	28		
RU/H	No pilot test	20	1	7	5	33	77	31.30
	Pilot test	10	2	21	11	44		
DRU	No pilot test	8	-	6	4	18	80	32.52
	Pilot test	15	8	30	9	62		
Master's L	No pilot test	3	-	4	2	9	42	17.07
	Pilot test	10	2	14	7	33		
Master's S	No pilot test	1	-	-	-	1	1	.41
	Pilot test	0	-	-	-	0		
Spec/Faith	No pilot test	0	-	-	-	0	4	1.63
	Pilot test	4	-	-	-	4		
Spec/Med	No pilot test	1	-	-	1	1	1	.41
	Pilot test	0	-	-	-	0		
Not found	No pilot test	2	-	-	2	2	2	.81
	Pilot test	0	-	-	-	0		

Reliability. Table 13 includes the frequencies of the types of methods used for the 114 authors who conducted reliability analysis of their survey data and reported results.

Table 13

*Frequencies of Reliability Methods (*n = *114)*

Technique	*n*	%
Internal consistency only	103	90.35
Split-half only	2	1.75
Test-retest only	1	.88
Multiple methods	5	4.39
Not specified	3	2.63

The multiple methods addressed by the authors included both internal consistency (using Cronbach's alpha) and either inter-rater reliability, factor analysis (n = 2), split-half, or test-retest. Three authors who reported coefficients did not specify the method used.

Of those coded as not conducting reliability analysis (n = 132), 50 authors did not mention reliability. Forty-one authors discussed reliability in some capacity; 23 were more detailed, but did not report results. Three authors discussed reliability for modified items, but not for their own. Five stated reliability evidence was not needed because the surveys were newly developed, and five discussed the reliability of the study or the methods used, not the reliability of the data obtained. Four mentioned reliability in general, and one reported reliability, but the analysis included modified items.

Reliability coefficients. The 114 authors reported reliability coefficients in combinations of only pilot test results, only main study results, or results of both; as well as for only total scales, only subscales, or results of both (see Table 14). Only two authors reported both total scale and subscale results from the pilot test and main study.

Table 14

*Reliability Reporting (*n = *114)*

Coefficients Reported	Pilot Test		Study Results		Both	
	n	%	*n*	%	*n*	%
Total scale only	20	17.54	24	21.05	7	6.14
Subscales only	5	4.39	31	27.20	3	2.63
Both	5	4.39	10	8.77	2	1.75

Two authors (1.75%) reported total scale reliability from the pilot test, and subscale reliability from the main study. Five authors (4.39%) reported total scale reliability for both the pilot test and main study, as well as the subscale reliability from the main study.

Of the authors who reported pilot test results (*n* = 49), 34 reported it only for the total scale and eight reported it only for the subscales. Seven authors reported both total scale and subscale results. The frequencies of coefficients for the 41 authors who reported total scales from the pilot tests are reported in Table 15. Two of the authors who reported total scale reliability for pilot test results included coefficients for multiple surveys. The coefficients reported in the dissertations were not rounded for inclusion in these ranges.

Table 15

Frequencies of Reliability Coefficients for Pilot Test Total Scales (n = 43)

Range	*n*	%
Strong: ≥ .80	31	72.09
Moderately strong: .60 to .79	8	18.60
Moderate: .40 to .59	3	6.98
Moderately weak: .20 to .39	0	-
Weak: < .20	1	2.33

Authors who included subscale coefficients for the pilot test results (n = 15) reported data for 73 subscales. The frequencies of coefficients for total study subscales is reported in Table 16.

Table 16

*Frequencies of Reliability Coefficients for Pilot Test Subscales (*n = *73)*

Range	*n*	%
Strong: ≥ .80	48	65.75
Moderately strong: .60 to .79	22	30.14
Moderate: .40 to .59	2	2.74
Moderately weak: .20 to .39	1	1.37
Weak: < .20	0	-

Of the authors who reported main study results (n = 84), 31 reported it only for the total scale and 36 reported it only for the subscales. Seventeen authors reported both total scale and subscale results. Six of the 48 authors who reported total scale reliability for main study results included coefficients for multiple surveys. The frequencies of reported coefficients for total scales is reported in Table 17.

Table 17

*Frequencies of Reliability Coefficients for Main Study Total Scales (*n = *55)*

Range	*n*	%
Strong: ≥ .80	39	70.91
Moderately strong: .60 to .79	14	25.45
Moderate: .40 to .59	2	3.64
Moderately weak: .20 to .39	0	-
Weak: < .20	0	-

Authors who included subscale coefficients for the main study results (n = 53) reported data for 237 subscales. The frequencies of coefficients for total study subscales is reported in Table 18.

Table 18

*Frequencies of Reliability Coefficients for Main Study Subscales (*n = *237)*

Range	*n*	%
Strong: ≥ .80	133	56.12
Moderately strong: .60 to .79	79	33.33
Moderate: .40 to .59	16	6.75
Moderately weak: .20 to .39	7	2.95
Weak: < .20	2	.84

Validity. The majority (n = 227) performed some type of validity analysis of their survey. Nineteen authors who did not conduct a pilot test also did not conduct a second method of validation; the majority performed both (n = 119). Table 18 includes the types of methods used and the frequencies of each.

Table 18

*Frequencies of Validity Methods (*n = *227)*

Technique	*n*	%
SME review only	131	57.71
Pilot test only	52	22.91
Factor analysis only	9	3.97
Alignment with literature only	8	3.52
Multiple methods	23	10.13
Not identified	4	1.76

The multiple methods addressed by the authors included combining SME review and factor analysis (n = 16), combining literature and SME reviews (n = 5), and combining literature and SME reviews with factor analysis (n = 2). Of those authors who did not discuss validity, nine mentioned the property but did not include evidence of analysis. Validity for these dissertations was discussed regarding validation of the study itself, not of the surveys.

Summary

Authors presented details of survey blueprints, pilot-testing the surveys, and evidence of reliability and validity, although those reporting any or all of these components varied widely. Of the 246 doctoral dissertations, the average score of item stem guidelines was 88.93%; the number of items either without accurate mechanics or that which were double-barreled may be a cause for concern. Characteristics of a variety of rating scales were assessed for descriptive purposes. Chapter five includes a discussion of these results.

CHAPTER FIVE

DISCUSSION

Reviewing published literature informs researchers and practitioners of current trends, new methodologies and findings, and current states of inquiry (Kieffer et al., 2001). The purpose of the current study was to identify trends in doctoral dissertations with student-developed surveys. "Focusing solely on dissertations also enables a clearer identification of methodological patterns in graduate student research" (Hallinger, 2011, p. 282).

Findings Related to the Literature

The results of data analysis indicated doctoral students who created their own surveys and rating scales for data collection were including pertinent descriptions in their chapters. However, not all students described, nor did some even acknowledge, necessary components for survey research. Findings from analysis of the dissertations are discussed and compared to relevant literature in the following sections.

Survey blueprint. Approximately 72% of the authors included some type of survey blueprint. These blueprints included definitions of the constructs to be measured by the survey or alignment of the survey or its items to the purpose of the study or research questions. Inclusion of this information is consistent with recommendations made in previous research (Coluci, 2012; Creswell, 2014; Dawis, 1987; Fink, 2003c, 2003d; Fowler, 1995, 2009; Ho, 2005; Netemeyer et al., 2003; Spector, 1992).

Item writing. Approximately 15% of the survey items did not have accurate mechanics. Although guidelines for academic writing (APA, 2010) were being used to assess the items that necessarily may not have been used by the authors, the mistakes found regarding grammar and punctuation were unwarranted. The use of accurate mechanics is a common recommendation for

writing items (Benson & Clark, 1982; Converse & Presser, 1986; Fink, 2003d; Fowler, 1992, 1995, 2009; Kitchenham & Pfleeger, 2002; Likert, 1932; Netemeyer et al., 2003; Spector, 2002). According to Fowler (2009), inadequate wording is to be avoided, as it affects the interpretations respondents make about the intent of the item. Inadequate wording results in inconsistent responses, thus affecting reliability of the data. The use of inadequate wording or inaccurate mechanics could affect a respondent's motivation to complete a survey, therefore affecting the response rate, or affect the respondent's opinion of the credibility of the researcher, all of which could affect the validity of the survey data.

Less than 2% of the items were of an inappropriate length. These were items that included definitions after the item stem was posed, which Fowler (1995) advises against doing. Including a definition after the item can affect validity, as the respondent may not understand the item posed if the definition is not clarified initially. The length of an item stem should be considered, as providing too much information can hinder responses (Benson & Clark, 1982; Converse & Presser, 1986).

The number of items that were double-barreled is of concern, not only because more than 25% of the items contained multiple concepts, but because of how these types of items can affect results, specifically reliability and validity. Multiple researchers advise against the use of double-barreled items (Converse & Presser, 1986; Fowler, 1995, 2009; Heiman, 2001; Likert, 1932; Netemeyer et al., 2003; Spector, 1992), as a true response cannot be gauged if the opinion of multiple concepts is being requested (Spector, 1992).

Converse and Presser (1986) recommended that items not be restrictive, as the use of negative terms in items stems can reduce reliability (Netemeyer et al., 2003; Spector, 1992); less than 3% of the items included negative terms. These items mainly included the use of "not",

which can be easily missed by respondents (Spector, 1992). It was determined all of the restrictive items in the sample could have been reworded in a positive manner.

Rating scales. The rating scales included a range of three to 13 response options. Spector (1992), Fowler (1995), Netemeyer et al., (2003), and Saris et al. (2010) recommended using up to 11 response options. One author went beyond this recommendation and included 13 response options, which is not recommended because no more meaningful information can be gathered from that many options (Fowler, 1995) and data gathered can impair reliability and validity (Netemeyer et al., 2003; Saris & Gallhofer, 2007). Of all the rating scales included in analysis, approximately 71% included the optimal number of five to seven response options (Fowler, 1995; Netemeyer et al., 2003).

Other characteristics of the rating scales collected for descriptive purposes included that the majority of the scales started with 1 versus 0. Schaeffer and Presser (2003) advised starting the scale with 0 can affect the quality of results. Only 4% of the rating scales were not continuous; approximately 2% of the surveys included scales that were both continuous and not continuous. Scales that are not continuous can affect the reliability of data, as the interpretation of the scales will differ across respondents (Fowler, 1995, 2009; Spector, 1992).

More of the surveys included a two-sided scale as opposed to a one-sided scale. Of those two-sided scales, the majority had the negative side of the scale presented first. This can affect the type of responses obtained, as respondents may be more likely to respond on the negative end of the scale (Fowler, 1995). Less than 6% of the scales did not have symmetrical response options on both sides, which affects reliability (Fink, 2003d; Fowler, 1995; Spector, 1992). Half of the rating scales were agree-disagree, in which Fowler (2009) states "researchers will have

more reliable, valid, and interpretable data if they avoid the agree-disagree question form" (p. 105), as this type of scale and its items are difficult to construct.

The scales were split equitably in regards to including anchors, or including both anchors and values. Of those surveys that included anchors, approximately 12% were not appropriate for the items posed. These response option anchors did not match the items posed (for example, a frequency response scale was provided for an agree-disagree item), which could lead to reliability and validity issues is the respondent is confused as to how to respond.

Approximately 12% of the surveys included scales in which not all points were labeled. For each of these, the scales' endpoints were labeled; in few cases the mid-point was also labeled with the endpoints, which has been shown to increase reliability (Fowler, 1995; Netemeyer et al., 2003). Not labeling all scale points can affect validity, as labeling each point of a scale adds further clarification for the respondents (Krosnick, 1999; Schaeffer & Presser, 2003).

The use of a neutral item was split almost evenly between being included and not being included. Recommendations on including this option vary, as it has been shown to increase validity (Krosnick, 1999) or both validity and reliability (Saris & Gallhofer, 2007). It has also been shown to not affect validity (Schaeffer & Presser, 2003).

Psychometric properties. Approximately 70% of the authors conducted a pilot study, which is recommended by multiple sources (Converse & Presser, 1986; Dawis, 1987; Draugalis et al., 2008; Fowler, 1995, 2009; Kitchenham & Pfleeger, 2002; Litwin, 2003; Netemeyer et al., 2003; Sanchez, 1992), as this component of survey research can improve the quality of the items and overall survey, as well as reduce measurement error (Fowler, 1995; Sanchez, 1992). Conducting a pilot study is a form of validating a survey (Netemeyer et al., 2003); therefore, dissertations that included the use of a pilot test was considered a form of validity, but other

types of validation were collected from the studies. More than 92% of the authors validated their surveys using a pilot test or some other method.

The values of the reliability coefficients for pilot study and main study total scales and subscales did not raise any flags, as more than 61% of the coefficients were .80 or greater. Few coefficients reported were below .60 (*Moderate* to *Weak*). Secondary pilot studies took place if the initial total scale and subscale coefficients were low. Of the seven total scale and subscale coefficients below .60 reported from the pilot studies, the authors of only two of the total scale results recalculated reliability using data from the main study, both of which showed improvement. Authors reported 27 total scale and subscale coefficients under .60 from main study data.

What is of concern was the number of authors who stated reliability analysis was not necessary to perform, or they were advised not to conduct it. A few stated it was not necessary because the survey was new. One author stated total scores were not being used, so reliability analysis was not necessary; however, the author was averaging item responses for statistical analysis. Some authors discussed reliability of the research design and methodology of the study, or of the process of sampling or administering the survey, as if it would replace description of the reliability of the data; or stated the survey was reliable because it was deemed valid in preliminary review.

There were few authors who reported no psychometric properties (pilot test, validity, or reliability). Although only approximately 7%, there was no evidence in these studies that the surveys were assessed to ensure they consistently measured what was purported to be measured. Of great concern is that more than half of the authors did not conduct reliability analysis. This finding is consistent with Thompson's (1994a) review of reliability practices in doctoral research

and Karadag's (2011) conclusion about the insufficient quality of psychometric reporting in doctoral dissertations.

The findings from the current study are also consistent with reviews of published literature. Reviews of articles revealed only up to 66% of authors reported both reliability and validity information (Bennett et al., 2011; Goodwin & Goodwin, 1985a; Qualls & Moss, 1996; Slaney et al., 2010; Slaney, Tkatchouk, Gabriel, & Maraun, 2009). A review of articles by Qualls and Moss (1996) revealed overall, approximately half of the instruments were lacking these pieces of psychometric properties; those who reported one piece were most likely reporting both. Both psychometrics were reported for 20% of the instruments; 49% of the instruments had either reliability or validity evidence reported. According to the researchers, most of the instruments were already established; a minority were new instruments, which were typically questionnaire surveys. The reported evidence for both established and new instruments was similar: approximately 34% of established and 31% of new instruments reported information on score reliability, and 22% of established and 16% of new instruments reported evidence of validity. The authors stated the *possibility* that researchers had, in fact, examined evidence of reliability and validity of their instruments but did not report the information. This occurred in the sample of dissertation in the current study: authors described reliability and validity, but did not report evidence of one or both. Hogan and Agnello (2004) called for at least minimal standards for reporting evidence of both psychometric properties (reliability and validity) in publications.

Reviews of published articles have found only up to 55% had reported validity (Goodwin & Goodwin, 1985b; Meier & Davis, 1990; Qualls & Moss, 1996; Tinsley & Irelan, 1989). In Slaney et al.'s (2009) review, approximately 87% only reported validity evidence. Similar

results were found in a subsequent study (Slaney et al., 2010). Hogan and Agnello's (2004) review of published psychological instruments showed that of the far fewer authors reporting validity (approximately 55%), 90% of those did so using correlations; few used factor analysis or other methods. Approximately 15% of the authors in the current study used some type of factor analysis for validity evidence, such as exploratory factor analysis, principal components analysis, or structural equation modeling.

The findings of the current study are mostly consistent with reliability reporting in published literature, in which only up to 52% of articles had reported reliability (Goodwin & Goodwin, 1985b; Green et al., 2011; Kieffer et al., 2001; Meier & Davis, 1990; Onwuegbuzie, 2002; Qualls & Moss, 1996; Slaney et al., 2009; Tinsley & Irelan, 1989; Willson, 1980). In Willson's (1980) study, up to 45% of the studies did not mention reliability at all. According to Onwuegbuzie (2002), "without information about the reliability of scores generated by the instruments utilized in the study, it is difficult to put a finding in its proper context" (p. 17).

The findings of this study, consistent with previous research, implicate that action is necessary to increase the psychometric reporting of surveys and any type of instrumentation used for data collection. Based on the current findings, recommendations for future research to improve the quality of dissertations and self-developed surveys through graduate training and other means are provided.

Recommendations for Future Research

Researchers are encouraged to expand on the findings of this study by continuing to review and assess the trends in survey research at the doctoral and professional levels. Focusing on the development, implementation, and evaluation component of the survey research, as well as the quality of the survey items and rating scales, used in these studies will further expand on

what is known about the practices of both students and professionals. Masters theses were not included in the sample for the current study; a review of these will further expand what is known about trends in graduate research.

Overall quality. Leech and Goodwin's (2008) assessment of doctoral programs showed that 80% offered a dissertation planning course, of which 65% of the programs required students to complete the course prior to their dissertation work. In the same sample, 28% had guidelines for assessing the quality of dissertations, 2% of which included rubrics. Recommendations to improve the overall quality of dissertations has come in the forms of the use of rubrics (Hamilton et al., 2010; Lovitts, 2007) and student peer reviews using a matrix sampling technique (Thompson, 1987). Knowing that a systematic review process will take place may encourage students to be meticulous in their attention to the quality of their dissertations. Burnett (1999) suggested the use of a Collaborative Cohort Model, which was shown in his study to aid students in acquiring a wider range of knowledge about other students' studies, research designs, and methodologies; aid students in the completion of their dissertations; and overall enhance the quality of those dissertations. From their sample of experienced dissertation examiners, Mullins and Kiley (2002) reported "mixed or confused theoretical and methodological perspectives" and "work that is not original" were characteristics of a poor dissertation, and the design and elegance were characteristic of an "outstanding" dissertation (p. 379).

According to Haller (1979), several factors can "contribute to students devaluing the intrinsic worth of their dissertations" (p. 58) such as programs not offering students opportunities to partake in smaller studies prior to beginning the dissertation process, or programs in which "research is not a central focus of discussion" (p. 58). Respondents in Porter et al.'s (1982) study also suggested "making use of postdoctoral appointments to polish research capabilities as

needed" (p. 481). Isaac et al. (1992) collected survey responses from faculty members regarding their perceptions and practices of the dissertation, comprehensive examinations, and overall doctoral process. The majority of faculty members (85%) indicated there was no alternative to the dissertation; the remaining members indicated the only viable alternative was in the form of already published research. Leech and Goodwin's 2008) study showed that 57% of the doctoral programs required students to engage in some type of research prior to the dissertation.

Instruments, checklists, rubrics, and the like have been created for the purpose of evaluating research, specifically focused on published literature. Such instruments have specified the quality of a study includes information that the research design addressed the research questions (Gottfredson, 1978; Suydam, 1968), instrumentation was reliable, valid, and suitable (Anderson & Kerr, 1968; Bennett et al., 2011; Johnson, 1957; Suydam, 1968), the statistical techniques were valid or appropriate (Anderson & Kerr, 1968; Gottfredson, 1978; Johnson, 1957; Suydam, 1968; Tuckman, 1990), and appropriate interpretations were made about the results (Anderson & Kerr, 1968; Gottfredson, 1978; Johnson, 1957; Suydam, 1968; Tuckman, 1990). It is recommended a global checklist or rubric be used to assess the quality of a doctoral dissertation, as well as to assess the characteristics of self-developed surveys.

Although not a purpose of the current study, but seemingly worthy to note, the writing and formatting of multiple dissertations was discouraging. There were grammatical and punctuation errors that ultimately should not occur in the writing of a project for a terminal degree. Fink (2003b) discussed reporting to a technical audience versus a nontechnical audience when reporting results of survey data; however, the points made also seemed appropriate for writing a doctoral dissertation. Do doctoral students write to a nontechnical audience, when they

should be reporting to a technical audience? A consideration of integrating scientific, academic, or technical writing into graduate curricula should be considered.

Graduate training. Researchers have contended that the lack of quality in dissertations may be attributed to graduate school training, noting deficiencies in program requirements. A convincing argument about an overarching theme in reviews of graduate programs was stated by Nyquist (2002): "individuals within and outside the academy today contend that the doctoral experience should better prepare students for their professional destinations more than it does" (p. 14). The training students receive in graduate school has been scrutinized not only for its quality, but for the lack of adequate preparation (Ponticell & Olivarez, 1997; Shultz et al., 1998): "Researchers who lack adequate research skills will not fully understand how to acquire and use data, when and how to use analyses, and how results pertain to practice" (Rossen & Oakland, 2008, p. 42).

Educators should be concerned about the quality of doctoral training, as there exists a connection between training in a graduate program and performance as a scholar. Goldstein (2012) alleged the methodological flaws in published literature could be attributed to the training the authors received while as students. The education and training students receive affects the research conducted by them while in school and as professionals, which in turn impacts their respective fields of study (Capraro & Thompson, 2008).

The importance of academic institutions continuously reviewing their dissertation requirements to improve both the process and the product has been stressed (Hamilton et al., 2010; Quarles & Roney, 1986), as "an adequate understanding and use of research design, measurement, statistical analysis, and other components of quantitative and qualitative science are needed to conduct and consume research" (Rossen & Oakland, 2008, p. 42). Graduate

training in research design and methodology; test construction and survey design; psychometric properties, specifically validity and reliability; and statistical analysis has received attention (Aiken et al., 2008; Aiken, West, Sechrest, & Reno, 1990; Capraro & Thompson, 2008; Felbinger et al., 1999; Hallinger, 2011; Hassad, 2010; Merenda, 1990, 1996, 2007; Nyquist, 2002).

Associating doctoral training to outcomes (e.g., dissertations) was not the purpose of the current study. Though not the focal point, this does warrant consideration due to the minimal research of the quality of doctoral dissertations. A comparison of graduate program curricula and research requirements to the programs' outcomes is a further recommendation for future research. Assessments within and between institutions, fields of study, or other characteristics may show that those institutions with minimal requirements in specific training is evident in the doctoral dissertations emanating from those programs.

Research design and methodology training. Aiken et al. (1990) and Aiken et al. (2008) surveyed the chairs of psychology departments in the United States and Canada that included a PhD program regarding the type and amount of training available to graduate students. "Students are encouraged to achieve a high level of research productivity without much consideration for training in the content, methodological, and quantitative skills necessary for good research" (Aiken et al., 1990, p. 729). The need to review program offerings in quantitative psychology was addressed due to training students in this area being "a matter of serious concern" (Aiken et al., 2008, p. 46). Aiken et al. (1990) concluded that graduate students need more training in methodology, especially since "inappropriate data analyses are still relatively common" (p. 729). The authors stressed that what was being taught in the graduate programs was not sufficient, arguing that their training, especially in advanced topics, is weak,

which could lead to flawed publications. According to Rossen and Oakland (2008), "no guidelines specify which courses should be required for students to achieve competency in research methodology" (p. 42).

Cone and Foster (1991) claimed that Aiken et al.'s (1990) view of a need for more measurement training was underemphasized, noting "psychology training persistently fails to recognize that measurement provides the foundation for all other scientific pursuits" (Cone & Foster, 1991, p. 653). Cone and Foster (1991) recommended graduate students should be required to "study the design, construction, and evaluation of dependent measures" and "expose students to classical measurement concepts" (p. 653). The lack of doctoral students in measurement fields, as well as measurement training, has continued to decline (Kline, 2011; Lambert, 1991; Merenda, 1990, 1996, 2007; Rossen & Oakland, 2008), evidenced by Leech and Goodwin (2008) finding that only 23% of their sample of programs required a measurement course. According to Whittington (1998),

> Further work is needed to explore how measurement is reported in other, less selective journals, in ERIC documents, and in other nonrefereed reports. The training of researchers, particularly those with academic goals, needs to be examined as well. ...I recommend further study of the kind and quality of measurement training graduate students receive in preparation for the doctorate, particularly the future "practitioners" of educational study. (p. 35)

In Capraro and Thompson's study (2008), approximately 72% of the sample of doctoral programs required completion of at least one quantitative methods course, and only approximately 46% required completion of a qualitative methods course. An astonishing 25.9% of the programs did not require completion of *either* quantitative or qualitative courses; however,

44.2% of the programs required completion of both types. None of the programs required completion of a mixed-methods course. Leech and Goodwin (2008) had similar results: 62% of the programs in their study required a quantitative research methods course; the same proportion required a qualitative research methods course. Only 22% required a mixed methods course. Goldstein's (2012) informal survey of 10 doctoral programs indicated only two required a qualitative course, and only three required a methods course beyond the student's master's degree program. Doctoral programs not requiring quantitative, qualitative, or mixed methods courses impacts the quality of research emanating from students and even practitioners. Quantitative and qualitative methods courses are needed if students plan to do research or "wish to teach in doctoral programs that prepare candidates who will engage in basic research" (Felbinger et al., 1999, p. 461).

Goldstein (2012) argued students overall were not receiving the training and experience needed to be "able to select, utilize, and apply such [fundamental] methods in sound, critical, and rigorous ways" (p. 495). He suggested programs change their research methods requirements, such as requiring both qualitative and quantitative courses and requiring more rigorous qualitative methods training (e.g., interviewing and observations). Specific to dissertation research, Goldstein (2012) recommended programs require students conduct research which would require taking a specialized or advanced methods course, and programs should not schedule research design courses until after the student has completed other methods courses. This type of structure would allow time and resources for a student to create a dissertation proposal, and to use the research design course as a vehicle to do so.

Faculty training. Aiken et al. (1990) concluded that when necessary, faculty should be retrained in the areas of methodology and quantitative analyses. This is supported by Hassad's

(2010) conclusion that faculty members were in need of training in order to serve on a dissertation committee. Noting the lack of quality in dissertations overall, Hassad (2010) isolated the lack of faculty members' knowledge of research design and statistics, also including the lack of support and mentoring in these areas to students. His qualitative study focused on the experiences of 25 students, faculty members, recent graduates, and statistical consultants cannot be generalized to all doctoral programs, but indicates there is concern that students are not receiving what is necessary to perform dissertation research. "By not having our current generation of doctoral students well trained in research methods, we run the risk of having future planning faculty inadequately equipped to train and mentor the next generation of doctoral students, with the cycle repeating" (Goldstein, 2012, p. 496). Rossen and Oakland (2008) emphasized that the decrease in faculty specializing in research methods could be associated with the decline in research methods preparation of graduate students.

Clay (2005) reported on the lack of students in quantitative psychology programs, noting that even mentors at the undergraduate level are not familiar with this subdiscipline. Continuing to review the research methods and measurement courses not only offered to graduate students, but *required* of graduate students, is recommended. It is also recommended programs continue to evaluate faculty members' knowledge of these concepts, as well as review requirements of research conducted by students prior to their dissertation work.

Survey research training. Designing, constructing, and administering a quality survey is essential to many pieces necessary for a study of the highest quality; the issues found in the sample of student-developed surveys was a lack of psychometric reporting. These self-developed surveys may have been created in a manner that was inadequate or inappropriate for their research purposes: do doctoral students not have the necessary knowledge or training

needed to create such a survey for their research? Concluded by Rossen and Oakland (2008), graduates will not have the necessary content knowledge of survey or questionnaire design when it's found that less than 20% of programs in their sample require the course. Developing a quality survey takes time, preparation, and resources (Alvarez et al., 2012; Benson & Clark, 1982; Coluci, 2012; Fecso, 1989; Netemeyer et al., 2003; Porter, 2011; Spector, 1992) to ensure the data obtained from the survey is reliable and valid, and the survey itself is appropriate for the sample and purpose for which it is being used. Creating a survey is

> a process in which aims and objectives are defined, the sample strategy is designed, a reliable and valid questionnaire is constructed, piloted and used to collect data. These data are then entered into software for analysis, analysis is made, and reports are formed and disseminated. (Alvarez et al., 2012, p. 2700)

A continued focus on graduate program offerings is recommended, specifically regarding survey research, because regrettably "it is notable that the vast majority of higher education programs do not offer courses on survey methodology" (Porter, 2011, p. 71). Aiken et al.'s (2008) review of graduate programs found that inclusion of survey research and test construction in the programs increased from 1990 to 2008; however, survey research was included in only 36% or less of the programs. Test construction was included in only 61% or less of the programs. Graduate students' competency of test construction was also addressed by Aiken et al. (2008). When addressing the percentage of students who could apply techniques to their own research, the percentage of psychology department chairs who indicated *most or all* of their students were competent with test construction was only 26%; those who indicated *few or none* of their students were competent with test construction was 33%. According to Spector (1992), there is not enough graduate instruction on how to construct rating scales, which are commonly

used in doctoral research. "Every psychologist who will make use of an existing scale or will invent a scale should have a fundamental measurement training" (Aiken et al., 2008, p. 47).

Orcher (2007) indicated students can create their own surveys without background or training in research methodology, and can evade reliability and validity analyses if creating the survey for a term project. According to Orcher (2007), a student can still create a "satisfactory" survey despite having limited time, resources, and knowledge of the process. Although most of the content of Orcher's (2007) book was comparable to other sources noted in the current study, and although the book is clearly not relevant to graduate studies, it is indicative of the issue that students – of any degree level – are not being properly trained on how to construct or administer quality surveys, or evaluate survey data. Researchers, authors, and reviewers of survey research need to "seriously consider the rigor that needs to be applied in the design, conduct, and reporting of survey research so that the reported findings…are a true contribution to the scientific literature" (Draugalis et al., 2008, p. 5).

Giving "rise to questions over validity of knowledge generated" (Alvarez et al., 2012, p. 2700), the use of online surveys gives a false impression that survey research is easy (Draugalis et al., 2008). With the increased use of online survey providers that offer users with templates to quickly create a survey (Creswell, 2014; Draugalis et al., 2008), easing "the design, construction and administration of e-surveys," (Alvarez et al., 2012, p. 2700), doctoral students may choose the survey research method of data collection, not only because it is easy, but because it is a cheap means of collecting large amounts of data that can be analyzed using descriptive or univariate statistics (Haller, 1979; Spector, 1992). "The quality of data from a survey is no better than the worst aspect of the methodology" (Fowler, 1995, p. 150). The surveys reviewed in the current study were used to assess content from a variety of areas. A review of the surveys

created for dissertation research in comparison to professional surveys already in use is warranted. Most authors in the current sample stated a survey did not exist for the purposes of their studies. Evaluating these claims would be worthy research.

Psychometric properties training. Students should be taught how to evaluate an instrument and its psychometric properties, and how to evaluate and report the properties of their own instruments (Meier, 1993). It was reported in Aiken et al.'s (2008) study of graduate programs that when addressing the percentage of students who could apply techniques to their own research, the percentage of psychology department chairs who indicated *most or all* of their students were competent with methods of reliability assessment increased from 27% in 1990 to 46% in 2008. Those who indicated *few or none* of their students were competent with reliability assessment decreased from 38% to 19%. The psychology department chairs who indicated *most or all* of their students were competent with using methods of validity assessment in their own research increased only from 22% in 1990 to 28% in 2008. Those who indicated *few or none* of their students were competent with validity assessment decreased from 44% to 31%. These values are discouraging, as even though the department chairs over time were more confident in their students' competencies of using methods of reliability and validity, still less than half of the students were identified as competent.

Meier and Davis (1990) stated when a study uses some type of scale, it is assumed that the instrument possesses "adequate psychometric properties, that is, that they reliably and validly assess the constructs in question" (p. 113). According to Thompson (2003), "Unfortunately, some people have difficulty addressing these essential psychometric issues, because their training has ill equipped them for this particular adventure" (p. 7) and "Regrettably, due to poor curricula in most doctoral programs...many doctoral students and university faculty do not

understand score reliability or what factors most affect score reliability" (Thompson, 2003, p. 19). This was demonstrated in the current study, in which some authors noted they were advised reliability analysis was not necessary to conduct. In Hogan et al.'s (2000) review of reliability practices, 6% did not report reliability for their instruments. The majority of those authors who did report reliability used coefficient alpha, a trend also seen in the current study. An additional recommendation is that graduate school coverage of coefficient alpha be increased, so students "gain a thorough understanding of this coefficient" (Hogan et al., 2000, p. 528).

In addition to their student-developed surveys, some authors in the current study also used established instruments to collect data and reported reliability coefficients from previous research. According to Thompson and Vacha-Haase (2000), "it becomes wasteful to allocate doctoral curriculum to teach students anything more sophisticated than how rotely to cite the unalterable reliability coefficients reported in a test manual" (p. 180). One-third of the journal articles reviewed in Vacha-Haase et al.'s (1999) study did not mention reliability, and "too many authors" reported coefficients from previous data (p. 340). Teaching students how to conduct reliability and validity analyses on their own data is essential, as "measurement textbooks on the whole do poorly at accurately communicating fundamental measurement concepts to our graduate students" (Thompson & Vacha-Haase, 2000, p. 175).

Evaluating the curricula requirements in the area of psychometrics is valuable, as in turn graduate programs may see better quality in dissertation research, similarly to how published research would be improved if researchers reported measures of reliability and validity (Tinsley & Irelan, 1989). "Greater attention to the reporting of psychometric properties...will increase researchers' awareness of the psychometric foundations of their work and of the need for innovation in scale development methodologies" (Meier & Davis, 1990, p. 115). Data collected

from a student-developed survey may be inadequate if the research design was poor, the survey was not accurately and systematically developed, or if the data was not subjected to validity and reliability analyses, thus culminating questionable results and interpretations. Future researchers are encouraged to assess the procedures used in doctoral and survey research. A review of reporting practices in dissertations, regardless of the type of instrumentation, will aid in further identifying trends of psychometric reporting.

Conclusion

"Doctoral dissertations do not represent the gold standard in the conduct of empirical research, and findings based solely on doctoral studies may not be generalized to the published literature" (Hallinger, 2011, p. 282). This type of thinking is discouraging, and should inspire educators and researchers to focus on graduate training so that doctoral research is no longer viewed as subpar. Dissertations can, in fact, become published in the form of an article in a refereed journal or as a book (Hamilton, 1993; Porter et al., 1982), as they are "a critical component of the knowledge creation endeavor" (Thompson, 1994a, p. 32) and "can be seen as reflecting the most current emphases in a research area" (Nelson & Coorough, 1994, p. 159). Publications derived from students' doctoral dissertations were found to be cited more often than other publications from the same authors (Porter et al., 1982). This leads to the conclusions that "dissertation research has real scientific merit" (Porter et al., 1982, p. 478).

Studies of measurement reporting in dissertations and published literature have shown inaccurate measurement reporting in research, regardless of the type of instrument used to collect data, and research on doctoral programs shows they are not parallel in curricula requirements, nor are some requiring essential training. It could be assumed that if students are not receiving proper education and training, they do not report measurement characteristics accurately in their

dissertations or fail to take the necessary steps in survey development. However, the majority of authors in the sample for this study were reporting key components of the development, implementation, and evaluation of their surveys. Further review of graduate school curricula and requirements may expose more current issues with the lack of training in topics specifically needed to conduct quality survey research, including research design and methodology; survey research; and psychometric reporting. The findings from this study should urge students and their committee members, as well as educators and practitioners, to be informed of best practices regarding survey research, as well as the importance of conducting quality research.

Regardless of the training, or lack thereof, provided to doctoral students, or the knowledge and experience of the students' dissertation committee members, the *quality* of any survey or dissertation should not be jeopardized. Students are responsible for ensuring their studies are quality research. Knowledgeable faculty members should be employed to aid doctoral students in their efforts to strive for quality output. Training students early in their professional careers *should* eliminate erroneous and inferior doctoral research, and expectantly extend to future published scholarly research of the highest quality, because "doctoral education is a noble endeavor that matters" (Nyquist, 2002, p. 20).

REFERENCES

Adams, G. B., & White, J. D. (1994). Dissertation research in public administration and cognate fields: An assessment of methods and quality. *Public Administration Review*, *54*(6), 565-576.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*(1), 32-50. doi:10.1037/0003-066X.63.1.32

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist*, *45*(6), 721-734.

Alidousti, S., Khosrowjerdi, M., Shahriari, P., Shirani, F., & Tarnoni, H. B. (2009). Designing a model for description of theses and dissertations information on a large scale. *Libri*, *59*, 180-197. doi:10.1515/libr.2009.017

Alvarez, J., Canduela, J., & Raeside, R. (2012). Knowledge creation and the use of secondary data. *Journal of Clinical Nursing*, *21*, 2699-2710. doi:10.1111/j.1365-2702.2012.04296.x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anderson, J., & Kerr, A. H. (1968). A checklist for evaluating educational research. *Educational Research*, *11*(1), 74-75.

Bailar, B. A., & Lanphier, C. M. (1978). A pilot study to develop survey methods to assess survey practices. *The American Statistician*, *32*(4), 130-132.

Barry, A. E., Chaney, B. H., Piazza-Gardner, A. K., & Chavarria, E. A. (2013). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*, 1-7. doi:10.1177/1090198113483139

Bennett, C., Khangura, S., Brehaut, J. C., Graham, I. D., Moher, D., Potter, B. K., & Grimshaw, J. M. (2011). Reporting guidelines for survey research: An analysis of published guidance and reporting practices. *PLoS Medicine*, *8*(8), 1-11. doi:10.1371/journal.pmed.1001069

Benson, J., & Clark, F. (1982). A guide for instrument development and validation. *The American Journal of Occupational Therapy*, *36*(12), 789-800.

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, *15*(4), 386-397. doi:10.1037/a0019626

Burnett, P. C. (1999). The supervision of doctoral dissertations using a Collaborative Cohort Model. *Counselor Education and Supervision*, *39*(1), 46-52.

Capraro, R. M., & Thompson, B. (2008). The educational researcher defined: What will future researchers be trained to do? *The Journal of Educational Research*, *101*(4), 247-253.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage Publications, Inc.

Clay, R. (2005). Too few in quantitative psychology: The subdiscipline sees shrinking numbers, but growing opportunities. *Monitor on Psychology*, *36*(8), 26-28.

Cleary, R. E. (1992). Revisiting the doctoral dissertation in public administration: An examination of the dissertations of 1990. *Public Administration Review*, *52*(1), 55-61.

Cleary, R. E. (2000). The public administration dissertation reexamined: An evaluation of the dissertations of 1998. *Public Administration Review*, *60*(5), 446-455.

Cohen, R. J., & Swerdlik, M. E. (2002). *Psychological testing and assessment: An introduction to test and measurement* (5th ed.). Boston, MA: McGraw-Hill Companies, Inc.

Coluci, M. Z. O. (2012). Measurement instruments for ergonomics surveys – methodological guidelines. In I. L. Nunes (Ed.), *Ergonomics: A systems approach* (pp. 119-130). Rijeka, Croatia: InTech.

Cone, J. D., & Foster, S. L. (1991). Training in measurement: Always the bridesmaid. *American Psychologist*, *46*(6), 653-654.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Thousand Oaks, CA: Sage Publications, Inc.

Coorough, C., & Nelson, J. (1997). The dissertation in education from 1950 to 1990. *Educational Research Quarterly*, *20*(4), 3-14.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Dane, F. C. (2011). *Evaluating research: Methodology for people who need to read research*. Thousand Oaks, CA: Sage Publications, Inc.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, *34*(4), 481-489.

Deming, W. E. (1947). Some criteria for judging the quality of surveys. *The Journal of Marketing*, *12*(2), 145-157.

Draugalis, J. R., Coons, S. J., & Plaza, C. M. (2008). Best practices for survey research reports: A synopsis for authors and reviewers. *American Journal of Pharmaceutical Education*, *72*(1), 1-6.

Fecso, R. (1989). What is survey quality: Back to the future. *Proceedings of the Section on Survey*. Retrieved from http://www.amstat.org/sections/srms/proceedings/papers/1989_013.pdf

Felbinger, C. L., Holzer, M., & White, J. D. (1999). The doctorate in public administration: Some unresolved questions and recommendations. *Public Administration Review*, *59*(5), 459-464.

Fincham, J. E., & Draugalis, J. R. (2013). The importance of survey research standards. *American Journal of Pharmaceutical Education*, *77*(1), 1-4.

Fink, A. (2003a). *How to design survey studies* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Fink, A. (2003b). *How to report on surveys* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Fink, A. (2003c). *How to sample in surveys* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Fink, A. (2003d). *The survey handbook* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Finney, S. J., & Pastor, D. A. (2012). Attracting students to the field of measurement. *Educational Measurement: Issues and Practice*, *31*(2), 2-7. doi:10.1111/j.1745-3992.2012.00228.x

Fowler, F. J., Jr. (1992). How unclear terms affect survey data. *Public Opinion Quarterly*, *56*, 218-231.

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Fowler, F. J., Jr. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Gliem, J. A., & Gliem, R. R. (2003, October). *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales*. Paper presented at the meeting of the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH. Retrieved from https://scholarworks.iupui.edu/bitstream/handle/1805/344/Gliem%20&%20Gliem.pdf?s..

Goldstein, H. A. (2012). The quality of planning scholarship and doctoral education. *Journal of Planning Education and Research*, *32*(4), 493-496. doi:10.1177/0739456X12449484

Goodwin, L. D., & Goodwin, W. L. (1985a). An analysis of statistical techniques used in the *Journal of Educational Psychology*, 1979-1983. *Educational Psychologist*, *20*(1), 13-21.

Goodwin, L. D., & Goodwin, W. L. (1985b). Statistical techniques in *AERJ* articles, 1979-1983: The preparation of graduate students to read the educational literature. *Educational Researcher*, *14*(2), 5-11.

Gottfredson, S. D. (1978). Evaluation psychological research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist*, *33*, 920-934.

Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in *Psychological Assessment*: Recognizing the people behind the data. *Psychological Assessment*, *23*(3), 656-669. doi:10.1037/a0023089

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861-871. doi:10.1093/poq/nfr057

Hall, B. W., Ward, A. W., & Comer, C. B. (1988). Published educational research: An empirical study of its quality. *Journal of Educational Research*, *81*(3), 182-189.

Haller, E. J. (1979). Questionnaires and the dissertation in educational administration. *Educational Administration Quarterly*, *15*(1), 47-66.

Hallinger, P. (2011). A review of three decades of doctoral studies using the Principal Instructional Management Rating Scale: A lens on methodological progress in educational leadership. *Educational Administration Quarterly*, *47*(2), 271-306. doi:10.1177/0013161X10383412

Hamilton, R. G. (1993, Summer). On the way to the professoriate: The dissertation. *New Directions for Teaching and Learning*, (54), 47-56.

Hamilton, P., Johnson, R., & Poudrier, C. (2010). Measuring educational quality by appraising theses and dissertations: Pitfalls and remedies. *Teaching in Higher Education*, *15*(5), 567-577. doi:10.1080/13562517.2010.491905

Harvey, L., & Green, D. (1993). Defining quality. *Assessment & Evaluation in Higher Education*, *18*(1), 9-34.

Hassad, R. A. (2010, July). Toward improving the quality of doctoral education: A focus on statistics, research methods, and dissertation supervision. In C. Reading (Ed.), *Data and context in statistics education: Towards and evidence-based society*. *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8). Ljubljana, Slovenia*. Voorburg, Netherlands: International Statistical Institute.

Hastings, J., & Stewart, J. (1983). An analysis of research studies in which "homemade" achievement instruments were utilized. *Journal of Research in Science Teaching*, *20*(7), 697-703.

Heiman, G. W. (2001). *Understanding research methods and statistics: An integrated introduction* (2nd ed.). Boston, MA: Houghton Mifflin Company.

Ho, F. W. H. (2005). Survey as a source of statistics and factors affecting the quality of survey statistics. *International Statistics Review*, *73*(2), 245-248.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, *64*(4), 802-812. doi:10.1177/0013164404264120

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*(4), 523-531.

Isaac, P. D., Quinlan, S. V., & Walker, M. M. (1992). Faculty perceptions of the doctoral dissertation. *Journal of Higher Education*, *63*(3), 241-268.

Johnson, G. B., Jr. (1957). A method for evaluating research articles in education. *Journal of Educational Research*, *51*(2), 149-151.

Kamler, B., & Thomson, P. (2008). The failure of dissertation advice books: Toward alternative pedagogies for doctoral writing. *Educational Researcher*, *37*(8), 507-514. doi:10.3102/0013189X08327390

Karadag, E. (2011). Instruments used in doctoral dissertations in educational sciences in Turkey: Quality of research and analytical errors. *Educational Sciences: Theory & Practice*, *11*(1), 330-334.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *The Journal of Experimental Education*, *69*(3), 280-309.

Kitchenham, B., & Pfleeger, S. L. (2002). Principles of survey research part 4: Questionnaire evaluation. *Software Engineering Notes*, *27*(3), 20-23. doi:10.1145/638574.638580

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Kohr, R. L., & Suydam, M. N. (1970). An instrument for evaluating survey research. *The Journal of Educational Research*, *64*(2), 78-85.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(5), 537-567.

Kupfersmid, J. (1988). Improving what is published. *American Psychologist*, *43*(8), 635-642.

Lambert, N. M. (1991). The crisis in measurement literacy in psychology and education. *Educational Psychologist*, *26*(1), 23-35.

Leech, N. L., & Goodwin, L. D. (2008). Building a methodological foundation: Doctoral-level methods courses in colleges of education. *Research in the Schools*, *15*(1), 1-8.

Lehman, J. W. (1974). Report on the ASA conference on surveys of human populations. *The American Statistician*, *28*(1), 30-34.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 5-55.

Litwin, M. S. (2003). *How to assess and interpret survey psychometrics* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Lovitts, B. E. (2007). *Making the implicit explicit: Creating performance expectations for the dissertation.* Sterling, VA: Stylus Publishing, LLC.

Lunenburg, F. C., & Irby, B. J. (2008). *Writing a successful thesis or dissertation: Tips and strategies for students in the social and behavioral sciences*. Thousand Oaks, CA: Corwin Press.

Meier, S. T. (1993). Revitalizing the measurement curriculum: Four approaches for emphasis in graduate education. *American Psychologist*, *48*(8), 886-891.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology*, *37*(1), 113-115.

Merenda, P. F. (1990). Brief note on graduate training in statistics, methodology, and measurement in psychology. *Perceptual and Motor Skills*, *71*, 1106.

Merenda, P. F. (1996). Note on the continuing decline of doctoral training in measurement. *Psychological Reports*, *78*, 321-322.

Merenda, P. F. (2007). Update on the decline in the education and training in psychological measurement and assessment. *Psychological Reports*, *101*, 153-155. doi:10.2466/PR0.101.1.153-155

Miller, P. V. (2010). Presidential address: The road to transparency in survey research. *Public Opinion Quarterly*, *74*(3), 602-606. doi:10.1093/poq/nfq038

Mullins, G., & Kiley, M. (2002). 'It's a PhD, not a Nobel Prize': How experienced examiners assess research theses. *Studies in Higher Education*, *27*(4), 369-386. doi:10.1080/0307507022000011507

Nelson, J. K., & Coorough, C. (1994). Content analysis of the PhD versus EdD dissertation. *Journal of Experimental Education*, *62*(2), 158-168. doi:10.1080/00220973.1994.9943837

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage Publications, Inc.

Newton, R. R., & Rudestam, K. E. (2013). *Your statistical consultant: Answers to your data analysis questions* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Nyquist, J. D. (2002). The PhD: A tapestry of change for the 21st century. *Change*, *34*(6), 12-20.

Onwuegbuzie, A. J. (2002). Common methodological, analytical, and interpretational errors in published educational studies: An analysis of the 1998 volume of the *British Journal of Educational Psychology*. *Educational Research Quarterly*, *26*(1), 11-22.

Orcher, L. T. (2007). *Conducting a survey: Techniques for a term project*. Glendale, CA: Pyrczak Publishing.

Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications in mental appraisal*. Boston, MA: Pearson Education, Inc.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications, Inc.

Perlmutter, D. D. (2006). Betrayed by your adviser. *Chronicle of Higher Education*, *52*(25), C3.

Ponticell, J. A., & Olivarez, A. (1997). Dissertation quality and Kerlinger's *methods myth*. *The Journal of Experimental Education*, *65*(2), 113-122.

Porter, A. L., Chubin, D. E., Rossini, F. A., Boeckmann, M. E., & Connolly, T. (1982). The role of the dissertation in scientific careers. *American Scientist*, *70*(5), 475-481.

Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education*, *35*(1), 45-76.

Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement*, *56*(2), 209-214.

Quarles, D. R., & Roney, R. K. (1986). Preparation, style, and format of doctoral dissertations in U.S. colleges and universities. *Research in Higher Education*, *25*(1), 97-108.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*(3), 306-322. doi:10.1037/1082-989X.11.3.306

Rossen, E., & Oakland, T. (2008). Graduate preparation in research methods: The current status of APA-accredited professional programs in psychology. *Training and Education in Psychology*, *2*(1), 42-49. doi:10.1037/1931-3918.2.1.42

Ruja, H. (1955). Citing references accurately. *American Psychologist*, *10*(7), 306-307.

Salkind, N. J. (2006). *Tests & measurement for people who (think they) hate tests & measurement*. Thousand Oaks, CA: Sage Publications, Inc.

Sanchez, M. E. (1992). Effects of questionnaire design on the quality of survey data. *Public Opinion Quarterly*, *56*(2), 206-217.

Saris, W. E., & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, *1*(1), 29-43.

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 61-79.

Saris, W. E., Van Wijk, T., & Scherpenzeel, A. (1998). Validity and reliability of subjective social indicators: The effect of different measures of association. *Social Indicators Research*, *45*, 173-199.

Schaeffer, N. C., & Dykema, J. (2011). Questions for surveys: Current trends and future directions. *Public Opinion Quarterly*, *75*(5), 909-961. doi:10.1093/poq/nfr048

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, *29*, 65-88. doi:10.1146/annurev.soc.29.110702.110112

Schuman, H. (1986). Ordinary questions, survey questions, and policy questions. *Public Opinion Quarterly*, *50*, 432-442.

Schuman, H., & Presser, S. (1980). Public opinion and public ignorance: The fine line between attitudes and nonattitudes. *American Journal of Sociology*, *85*(5), 1214-1225.

Shultz, K. S., Riggs, M. L., & Kottke, J. L. (1998). The need for an evolving concept of validity in industrial and personnel psychology: Psychometric, legal, and emerging issues. *Current Psychology*, *17*(4), 265-286.

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., Ferguson, L. P., Knudsen, J. R. S., & Legere, J. C. (2010). A review of psychometric assessment and reporting practices: An examination of measurement-oriented versus non-measurement-oriented domains. *Canadian Journal of School Psychology*, *25*(3), 246-259. doi:10.1177/0829573510375549

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., & Maraun, M. D. (2009). Psychometric assessment and reporting practices: Incongruence between theory and practice. *Journal of Psychoeducational Assessment*, *27*(6), 465-476. doi:10.1177/0734282909335781

Snyder, P. (2000). Guidelines for reporting results of group quantitative investigations. *Journal of Early Intervention*, *23*(3), 145-150.

Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Newbury Park, CA: Sage Publications, Inc.

Spriestersbach, D. C., & Henry, L. D., Jr. (1978). The Ph.D. dissertation: Servant or master? *Improving College and University Teaching*, *26*(1), 52-55, 60.

Suydam, M. N. (1968). An instrument for evaluating experimental educational research reports. *The Journal of Educational Research*, *61*(5), 200-203.

Tansey, T. N., Zanskas, S. A., & Phillips, B. N. (2012). Doctoral dissertation research in rehabilitation counseling: 2005-2007. *Rehabilitation Counseling Bulletin*, *55*(2), 103-125. doi:10.1177/0034355211431969

Tewari, D. D. (2012). Examination of doctoral theses/dissertations: Models, practices, and guidelines. *African Journal of Business Management*, *6*(9), 3438-3448. doi:10.5897/AJBM11.652

Thompson, B. (1987, January). *Peer review of doctoral dissertations as a quality control mechanism: Some methods and examples*. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Thompson, B. (1988, November). *Common methodology mistakes in dissertations: Improving dissertation quality*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY.

Thompson, B. (1994a, April). *Common methodology mistakes in dissertations, revisited*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Thompson, B. (1994b). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Thompson, B. (1994c). Inappropriate statistical procedures in counseling research: Three pointers for readers of research literature. *ERIC Digest*. (EDO-CG-95-33)

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 3-23). Thousand Oaks, CA: Sage Publications, Inc.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *Journal of Counseling & Development* research articles. *Journal of Counseling & Development*, *76*, 436-441.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*(2), 174-195.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Tinsley, D. J., & Irelan, T. M. (1989). Instruments used in college student affairs research: An analysis of the measurement base of a young profession. *Journal of College Student Development*, *30*, 440-447.

Tuckman, B. W. (1990). A proposal for improving the quality of published educational research. *Educational Researcher*, *19*(9), 22-25.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*(4), 562-569.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education*, *67*(4), 335-341.

Vagias, W. M. (2006). *Likert-type scale response anchors*. Clemson University: Clemson International Institute for Tourism & Research Development.

Vockell, E. L., & Asher, W. (1974). Perceptions of document quality and use by educational decision makers and researchers. *American Educational Research Journal*, *11*(3), 249-258.

Ward, A. W., Hall, B. W., & Schramm, C. F. (1975). Evaluation of published educational research: A national survey. *American Educational Research Journal*, *12*(2), 109-128.

West, C. K., Carmody, C., & Stallings, W. M. (1983). The quality of research articles in the *Journal of Educational Research*, 1970 and 1980. *Journal of Educational Research*, *77*(2), 70-76.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, *58*(1), 21-37.

Wick, J. W., & Dirkes, C. (1973). Characteristics of current doctoral dissertations in education. *Educational Researcher*, *2*(7), 20-21.

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and Explanations. *American Psychologist*, *54*(8), 594-604.

Willson, V. L. (1980). Research techniques in *AERJ* articles: 1969 to 1978. *Educational Researcher*, *9*(6), 5-10.

Winter, R., Griffiths, M., & Green, K. (2000). The 'academic' qualities of practice: What are the criteria for a practice-based PhD? *Studies in Higher Education*, *25*(1), 25-37. doi:10.1080/030750700115993

Appendices

Appendix A: References for Survey Research Components

Component	References
Survey Blueprint	Bennett et al., 2011; Benson & Clark, 1982; Coluci, 2012; Converse & Presser, 1986; Creswell, 2014; Dawis, 1987; Fink, 2003b, 2003d; Fowler, 2009; Ho, 2005; Kohr & Suydam, 1970; Netemeyer et al., 2003; Spector, 1992
Pilot test	Benson & Clark, 1982; Coluci, 2012; Converse & Presser, 1986; Creswell, 2014; Draugalis et al., 2008; Fowler, 1992, 1995, 2009; Hastings & Stewart, 1983; Ho, 2005; Kitchenham & Pfleeger, 2002; Litwin, 2003; Netemeyer et al., 2003; Sanchez, 1992; Spector, 1992
Reliability	Anderson & Kerr, 1968; Barry et al., 2013; Bennett et al., 2011; Benson & Clark, 1982; Carmines & Zeller, 1979; Coluci, 2012; Converse & Presser, 1986; Creswell, 2014; Dawis, 1987; Draugalis et al., 2008; Fink, 2003b, 2003d; Fowler, 1995, 2009; Gliem & Gliem, 2003; Hallinger, 2011; Hastings & Stewart, 1983; Heiman, 2001; Kieffer et al., 2001; Kline, 2001; Kohr & Suydam, 1970; Litwin, 2003; Lunenburg & Irby, 2008; Meier & Davis, 1990; Netemeyer et al., 2003; Qualls & Moss, 1996; Shultz et al., 1998; Snyder, 2000; Spector, 1992; Thompson, 1988, 1994a, 1994b, 1994c, 2000, 2003; Thompson & Vacha-Haase, 2000; Tinsley & Irelan, 1989; Tuckman, 1990; Ward et al., 1975; Whittington, 1998; Wilkinson & the APA Task Force on Statistical Inference (1999)
Validity	Alvarez et al., 2012; Anderson & Kerr, 1968; Barry et al., 2013; Bennett et al., 2011; Benson & Clark, 1982; Carmines & Zeller, 1979; Coluci, 2012; Converse & Presser, 1986; Creswell, 2014; Draugalis et al., 2008; Fink, 2003b, 2003d; Fowler, 1995, 2009; Hastings & Stewart, 1983; Heiman, 2001; Hogan & Agnello, 2004; Kline, 2011; Kohr & Suydam, 1970; Litwin, 2003; Lunenburg & Irby, 2008; Meier & Davis, 1990; Netemeyer et al., 2003; Qualls & Moss, 1996; Shultz et al., 1998; Spector, 1992; Thompson, 2003; Tinsley & Irelan, 1989; Tuckman, 1990; Ward et al., 1975; Whittington, 1998; Wilkinson & the APA Task Force on Statistical Inference (1999)

Appendix B: Coding Schema for Item Writing and Rating Scale Guidelines

Component	Guideline	Description	Code
Stem	Accurate mechanics	Correct syntax, punctuation, and other mechanics according to APA (2010).	0 = No mistakes made 1 = Mistake(s) made
	Appropriate length	Introduction of a definition or description of a concept occurs prior to the question or statement posed.	0 = Not appropriate 1 = Appropriate N/A = Not applicable
	Double-barreled	Question or statement about a single concept (is mutually exclusive or does not include multiple concepts).	0 = Not double-barreled 1 = Double-barreled
	Positively worded	Exclusion of negative words or terms (i.e., *except*, *not*).	0 = Negatively worded 1 = Positively worded
Rating Scale	Scale starts with 0 or 1	The rating scale begins with a 0 or 1 (if values are used).	0 = Starts with 0 1 = Starts with 1 N/A = Not applicable (no values used in scale)
	Continuous scale	The scale and its response options are along a continuum; options are not out of order or missing.	0 = No 1 = Yes
	One- or two-sided scale	A one-sided scale is used (positive [zero to positive values] or negative [negative values to zero]), or a two-sided scale is used (negative to positive values).	0 = One-sided negative 1 = One-sided positive 2 = Two-sided
	Side of scale presented first	Positive or negative side of two-sided scale presented first.	0 = Negative 1 = Positive N/A = Not applicable
	Symmetrical	Options on both sides of the midpoint of the two-sided scale are symmetrical (i.e., match across the scale).	0 = Not symmetrical 1 = Symmetrical N/A = Not applicable
	Number of options	Number of response options provided in the scale.	Value
	Type of options	Anchor, value, or both presented.	0 = Anchor 1 = Value 2 = Both
	Anchors are appropriate	Anchor wording matches stem wording; the anchors are continuous across the scale (i.e., same adjectives).	0 = Not appropriate 1 = Appropriate N/A = Not applicable (no labels used)
	Each scale point is labeled	Each scale point is labeled with an anchor or value. If both are used, they are not contradictory.	0 = No 1 = Yes
	Neutral option	Inclusion of a neutral response option (e.g., *Don't Know*, *Neither Agree nor Disagree, Neutral, No Opinion*).	0 = No 1 = Yes

Appendix C: Coding Schema for Pilot Test, Reliability, and Validity Methods

Component	Method	Description
Pilot Study		If the author stated a pilot test was conducted, the dissertation was coded as completing a pilot test.
	Sample	Author only used a representative sample for the pilot test.
	SME review	Author only used a subject matter expert review for the pilot test.
	Both	Author used both a representative sample and a subject matter expert review for the pilot test.
	Not identified	Author indicated a pilot test was conducted, but did not specify the method.
Reliability		If the author stated reliability was conducted *and* results were reported, the dissertation was coded as reliability analysis completed.
	Internal consistency	Author only used internal consistency as the reliability method.
	Split-half	Author only used split-half as the reliability method.
	Test-retest	Author only used test-retest as the reliability method.
	Multiple methods	Author used a combination of internal consistency and inter-rater reliability, factor analysis, split-half, or test-retest as methods of reliability.
Validity		If the author stated a method was used to validate the survey *beyond* conducting a pilot test, *and* included evidence of analysis, the dissertation was coded as validity analysis completed.
	SME review only	Author only used a subject matter expert review for validation.
	Pilot test only	Author specified the pilot test was used for validation. No other method specified.
	Factor analysis only	Author only used factor analysis for validation.
	Alignment with literature only	Author only aligned the survey with existing literature for validation.
	Combination	Author used a combination of subject matter expert review, literature review, and factor analysis for validation.
	Not identified	Author indicated validity was conducted, but did not specify the method.